SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 1-8, February 1994

001

EXPLICIT CONSTRUCTIONS OF DEPTH-2 MAJORITY CIRCUITS FOR COMPARISON AND ADDITION*

NOGA ALON† AND JEHOSHUA BRUCK‡

Abstract. All Boolean variables here range over the two-element set $\{-1, 1\}$. Given n Boolean variables $x_1, \ldots, x_n$, a nonmonotone MAJORITY gate (in the variables x_i) is a Boolean function whose value is the sign of $\sum_{i=1}^{n} \varepsilon_i x_i$, where each ε_i is either 1 or -1. The COMPARISON function is the Boolean function of two n-bits integers X and Y whose value is -1 if and only if $X \geqq Y$. An explicit sparse polynomial whose sign computes this function is constructed. Similar polynomials are constructed for computing all the bits of the summation of the two numbers X and Y. This supplies explicit constructions of depth-2 polynomial-size circuits computing these functions, which use only nonmonotone MAJORITY gates. These constructions are optimal in terms of the depth and can be used to obtain the best-known explicit constructions of MAJORITY circuits for other functions like the product of two n-bit numbers and the maximum of n n-bit numbers. A crucial ingredient is the construction of a discrete version of a sparse "delta polynomial"—one that has a large absolute value for a single assignment and extremely small absolute values for all other assignments.

Key words. threshold function, majority circuits, comparison function, addition function, error-correcting codes

AMS subject classifications. 68Q15, 68RXX

1. Introduction.

In this paper, we address the problem of computing the COMPARISON and ADDITION functions of two n-bit numbers using circuits of (nonmonotone) MAJORITY gates. Throughout this paper, a Boolean function will be defined as $f: \{1, -1\}^n \rightarrow \{1, -1\}$, namely, logical 0 and logical 1 are represented by 1 and -1, respectively.

We first define a few concepts.

1.1. Definitions.

DEFINITION 1. A linear threshold function $f(X)$ is a Boolean function such that

$$f(X) = \operatorname{sgn}(F(X)) = \begin{cases} 1 & \text{if } F(X) \geqq 0, \\ -1 & \text{if } F(X) < 0, \end{cases}$$

where

$$F(X) = w_0 + \sum_{i=1}^{n} w_i x_i .$$

The coefficients w_i are called the *weights* of the threshold function. We denote the class of linear threshold functions by LT_1. Note that the weights can be arbitrary real numbers. It is more interesting to consider the subclass of LT_1, which we call $\widehat{LT}_1$ of functions that can be written with "small" weights. Each function

$$f(X) = \operatorname{sgn}\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$$

* Received by the editors August 26, 1991; accepted for publication (in revised form) October 2, 1992.

† IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120-6099 and Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. This author's research was supported in part by the Fund for Basic Research administered by the Israel Academy of Sciences.

‡ IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120-6099.

in $\widehat{LT}_1$ is characterized by the property that the weights w_i are integers bounded by a polynomial in n, namely, $|w_i| \leqq n^c$ for some constant $c > 0$.

Note that, when we say that a Boolean function belongs to a certain complexity class (like LT_1), we actually mean that the family of Boolean functions (as defined for all n) belong to that class.

In this paper, we will be mostly interested in linear threshold functions in which the weights are either 1 or -1. Clearly, the elements that compute those functions are nonmonotone analogues of usual MAJORITY gates, which we call here, for short, MAJ gates.

DEFINITION 2. An MAJ gate computes a linear threshold function with weights that are either 1 or -1.

We are interested in circuits that consist of MAJ gates. Define MAJ_k to be the class of Boolean functions that can be computed by a polynomial-size depth-k circuit of MAJ gates, where the depth of the circuit is the number of gates on the longest path from the input to the output. Note that MAJ_k is equivalent to the class $\widehat{LT}_k$, which is the class of Boolean functions that can be computed by depth-k polynomial-size circuits of linear threshold elements with polynomial weights.

After presenting the computational model, let us introduce the functions that we would like to compute.

Let $X = (x_n, x_{n-1}, \ldots, x_1)$ and $Y = (y_n, y_{n-1}, \ldots, y_1)$ be two vectors in $\{1, -1\}^n$. Let a and b be the integers that correspond to X and Y, respectively. Since our convention is that a logical 0 is represented by 1 and a logical 1 is represented by -1, this means that $a = \sum_{i=1}^n (1 - x_i)2^{i-2}$ and $b = \sum_{i=1}^n (1 - y_i)2^{i-2}$.

DEFINITION 3. The COMPARISON function $C(X, Y)$ is the Boolean function that is -1 if and only if $a \geqq b$.

DEFINITION 4. Let $c = a + b$ and let $Z = (z_{n+1}, z_n, \ldots, z_1)$ be the binary representation of c. Then the ADDITION function is ADD $(X, Y) = Z$.

1.2. Motivation and known results. Why is it interesting to consider these two functions?

1) It was proved in [15] that the PRODUCT of two n-bit numbers is in MAJ_4. However, the proof is nonconstructive. Our construction for the ADDITION function can be used to explicitly describe a depth-4 MAJ circuit for PRODUCT. A different way of obtaining such an explicit depth-4 circuit has been recently found independently in [10]. We note here that depth 3 is the lower bound [11].

2) It was proved in [15] that any LT_1 function (one that can have large weights) is in MAJ_3. This proof is also nonconstructive. Our construction for COMPARISON can be used to explicitly construct a depth-3 MAJ circuit for any LT_1 function.

3) The construction for COMPARISON can be used also as a building block for a depth-3 circuit that sorts n n-bits numbers (see [8], [16], [17]).

It is known [8], [17] that COMPARISON $\in \text{MAJ}_3$ and ADDITION $\in \text{MAJ}_4$. It was also observed in [15] that COMPARISON $\in LT_1$, namely, the COMPARISON function can be computed by a single linear threshhold element. However, this linear threshold element has exponentially big weights. As shown in [15], COMPARISON $\notin \widehat{LT}_1$. On the other hand, using the results in [7], it was proved in [15] that both COMPARISON and ADDITION are in MAJ_2. The proofs in [15] are existence proofs, while finding explicit constructions was left as an open problem, which we solve here.

1.3. The main contribution. Our main contributions in this paper are explicit constructions of depth-2 polynomial size circuits of MAJ gates that compute the COMPARISON and ADDITION functions. Actually, we show that the COMPARISON and AD-

DITION functions can be computed as sign functions of explicit sparse polynomials (i.e., polynomials with $n^{O(1)}$ monomials and with $1, -1$-coefficients). In [6] it is proved that any function that can be computed as a sign of such a polynomial is also in MAJ_2. Hence, COMPARISON and ADDITION are in MAJ_2. The key to the construction is the idea that we can construct sparse polynomials that have the property of a "discrete delta function" in the sense that the value of the polynomial is very large for X being the all-1 vector and extremely small for all other values. The construction of these polynomials, which we call delta polynomials, is presented in the next section. In §3 we use the delta polynomials as a building block in the construction of depth-2 MAJ circuits for COMPARISON and ADDITION. These constructions can be practical, as they may be used in the actual design of small depth circuits for addition and multiplication based on MAJ gates. Section 4 contains some concluding remarks, extensions, and open problems.

2. Character sums, linear codes, and delta polynomials. Let $x_1, \ldots, x_n$ be n variables, where each x_i ranges over the two-element set $\{-1, 1\}$. Since $x_i^2 = 1$ for all i, every polynomial in the variables x_i can be represented as a multilinear polynomial. We thus define a *monomial* in the variables x_i to be a product of a subset of the set of variables with a coefficient $+1$ or -1, i.e., a product of the form $\varepsilon_j \prod_{i \in A} x_i$, where $\varepsilon_j \in \{-1, 1\}$ and $A \subseteq \{1, \ldots, n\}$.

A polynomial in the variables x_i, above, is called t-sparse if it is the sum of at most t monomials. We are mainly interested in the case that t is at most $n^{O(1)}$.

For a vector $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_n\}$, where $\varepsilon_i \in \{-1, 1\}$, and for a positive real c, we call a polynomial $P(x_1, \ldots, x_n)$ a *delta polynomial* for ε and c if there are two positive constants a and b satisfying $a/b \geqq c$ such that (i) $P(\varepsilon_1, \ldots, \varepsilon_n) = a$ and (ii) for all $(x_1, \ldots, x_n) \in \{-1, 1\}^n$, which satisfies $(x_1, \ldots, x_n) \neq \varepsilon$, $|P(x_1, \ldots, x_n)| \leqq b$.

Therefore, P is a delta polynomial for ε and c if it attains a positive value at ε and the absolute value of P on any other point in $\{-1, 1\}^n$ is smaller by at least a factor of c.

Observe that the polynomial $\prod_{i=1}^{n} (1 + x_i)$ is a delta polynomial for $(1, 1, \ldots, 1)$ and any positive c. However, this polynomial is a sum of exponentially many monomials. Our objective in this section is to explicitly construct relatively sparse delta polynomials. A probabilistic construction follows from the techniques presented in [7], [15]; however, their explicit construction seems to be more difficult.

We can easily check that, if $P(x_1, \ldots, x_n)$ is a delta polynomial for $(1, 1, \ldots, 1)$ and c, then, for any vector $(\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$, $P(\varepsilon_1 x_1, \ldots, \varepsilon_n x_n)$ is a delta polynomial for ε and c, which has exactly the same number of monomials as P. Thus we may restrict our attention to the construction of sparse delta polynomials for $(1, 1, \ldots, 1)$.

Our construction can be obtained by using linear error-correcting codes over $GF(2)$ that have certain distance properties. We discuss this general approach at the end of the section. Now we present in more detail one such construction, which is based on the properties of the quadratic residue character that are proved using Weil's famous theorem, known as the Riemann hypothesis for curves over finite fields [18]. These properties have been used before to derive the pseudorandom properties of Paley graphs and quadratic tournaments [9] (see also [5], [2]) and have also been used in the analysis of certain randomized algorithms for various number-theoretic problems (see [4], [13]). Other constructions can be given based on some of the ideas of [1] and [12], together with the known constructions of expander-graphs or based on the results of [3]. For our purposes, the quadratic residue construction suffices, and we thus only describe this construction in detail and comment briefly at the end of the section on how to obtain additional similar constructions.

Let q be an odd prime power and let χ be the quadratic residue character defined on the elements of the finite field $GF(q)$ by $\chi(y) = y^{(q-1)/2}$. Equivalently, $\chi(y)$ is 1 if y is a nonzero square, 0 if y is 0, and -1, otherwise. Suppose that $q \geqq n$ and let $B = \{b_1, \dots, b_n\}$ be an arbitrary subset of cardinality n of $GF(q)$. Consider the following polynomial in the n variables $x_1, \dots, x_n$:

$$P_B(x_1, \dots, x_n) = \sum_{y \in GF(q)\setminus B} \prod_{i=1}^{n} \frac{1 + \chi(y - b_i) + x_i(1 - \chi(y - b_i))}{2}.$$

Observe that P_B is a sum of exactly $q - n$ monomials, since, for each fixed y in $GF(q)\setminus B$, the quantity $(1 + \chi(y - b_i) + x_i(1 - \chi(y - b_i)))/2$ is either 1 or x_i.

THEOREM 1. *For every odd prime power q and for every subset B of cardinality n of $GF(q)$, the polynomial P_B defined above satisfies*

(i) *$P_B(1, 1, \dots, 1) = q - n$ and*

(ii) *For every $(x_1, \dots, x_n) \in \{-1, 1\}^n$ that is not $(1, 1, \dots, 1)$, $|P_B(x_1, \dots, x_n)| \leqq (n - 1)q^{1/2}$.*

*Therefore, P_B is a $(q - n)$-*sparse *delta polynomial for $(1, 1, \dots, 1)$ and $c = (q - n)/(n - 1)q^{1/2}$.*

Note that, when q is a prime and B is simply the set $\{1, 2, \dots, n\}$, the expression for the polynomial P_B is relatively simple.

To prove Theorem 1, we need the following known estimate for character sums, due to Weil [18] (see also [14, Thm. 2C, p. 43]; the lemma stated below is a special case).

LEMMA 1. *Let q be an odd prime power and let F be the field $GF(q)$. Let $f(y)$ be a nonconstant polynomial over F that decomposes into the product of m distinct linear factors. Then, for the quadratic character χ,*

$$\left| \sum_{y \in F} \chi(f(y)) \right| \leqq (m - 1)q^{1/2}.$$

Proof of Theorem 1. Since P_B is a sum of $q - n$ monomials, it is $(q - n)$-sparse. Moreover, since the coefficient of every monomial is 1, it follows that $P_B(1, 1, \dots, 1) = q - n$. Suppose now that $(x_1, \dots, x_n) \neq (1, 1, \dots, 1)$ is a vector in $\{-1, 1\}^n$. Put $I = \{i : 1 \leqq i \leqq n, x_i = -1\}$, $J = \{b_i : i \in I\}$. By substituting the values of the variables x_i and by the fact that the quadratic character is multiplicative, we obtain

$$\prod_{i=1}^{n} \frac{1 + \chi(y - b_i) + x_i(1 - \chi(y - b_i))}{2} = \prod_{i \in I} \chi(y - b_i) = \chi\left(\prod_{i \in I} (y - b_i) \right).$$

Define $f(y) = \prod_{i \in I} (y - b_i)$. Observe that, for the quadratic character χ, $\chi(f(y)) = 0$ whenever y is equal to one of the elements b_i for $i \in I$. Therefore,

$$\begin{aligned} P_B(x_1, \dots, x_n) &= \sum_{y \in GF(q)\setminus B} \prod_{i=1}^{n} \frac{1 + \chi(y - b_i) + x_i(1 - \chi(y - b_i))}{2} \\ &= \sum_{y \in GF(q)\setminus B} \chi(f(y)) \\ &= \sum_{y \in GF(q)\setminus (B\setminus J)} \chi(f(y)) \\ &= \sum_{y \in GF(q)} \chi(f(y)) - \sum_{y \in B\setminus J} \chi(f(y)). \end{aligned}$$

Observe that, since I is not empty and since the elements b_i are distinct, we can apply Lemma 1 to $f(y)$ and obtain, by the triangle inequality,

$$|P_B(x_1, \ldots, x_n)| \leqq \left| \sum_{y \in GF(q)} \chi(f(y)) \right| + \left| \sum_{y \in B \setminus J} \chi(f(y)) \right| \leqq (|I| - 1)q^{1/2} + n - |I|.$$

The quantity $(|I| - 1)q^{1/2} + n - |I|$ is clearly an increasing function of $|I|$, and, since $|I| \leqq n$, this quantity is at most $(n-1)q^{1/2}$. This completes the proof. □

Linear codes and delta polynomials. The above argument can be modified to obtain a similar construction of a delta polynomial from any linear error-correcting code over $GF(2)$ with length that is polynomial in the dimension and with the property that the Hamming weight of any nonzero codeword is sufficiently close to half the length. Here is a sketch of the argument. Let $A = (a_{ij})_{1 \leqq i \leqq n, 1 \leqq j \leqq t}$ be the generating 0, 1-matrix of a linear error-correcting code of length t and dimension n and suppose that the Hamming weight of each nonzero codeword is between $(1-\varepsilon)(t/2)$ and $(1+\varepsilon)(t/2)$. Let $P_A = P_A(x_1, \ldots, x_n)$ be the polynomial defined by

$$P_A(x_1, \ldots, x_n) = \sum_{j=1}^{t} \prod_{i; a_{ij} = 1} x_i.$$

Clearly, $P_A(1, \ldots, 1) = t$, and it is not difficult to check that, for every $(x_1, \ldots, x_n) \in \{-1, 1\}^n$ that is not $(1, \ldots, 1)$,

$$|P_A(x_1, \ldots, x_n)| \leqq \varepsilon t,$$

since $P_A(x_1, \ldots, x_n)$ is precisely the difference between the number of 0's and the number of 1's in the codeword defined by the sum (in $GF(2)$) of all rows i of A such that $x_i = -1$.

The polynomial P_B described in Theorem 1 is a special case of the above construction, which corresponds to a linear error-correcting code of dimension n and length $q - n$, in which the Hamming weight of any nonzero codeword is between $(q-n)/2 - (n-1)q^{1/2}/2$ and $(q-n)/2 + (n-1)q^{1/2}/2$. Three additional simple constructions of linear codes whose parameters are asymptotically comparable to this one (up to some polylogarithmic factors) are given in [3], and any of these can be used for constructing a sparse delta polynomial in the manner described above.

3. The constructions. In this section, we prove that the COMPARISON and ADDITION functions can be computed as sign functions of (explicit) sparse polynomials. From a (simple) result in [6], this implies that both functions can be computed by an explicit depth-2 polynomial-size circuit of MAJ elements. Both constructions apply the delta polynomials described in the previous section.

First, we note that the following is an equivalent description of the COMPARISON function: For $X, Y \in \{1, -1\}^n$, $C(X, Y) = -1$ if and only if either $X = Y$ or there exists an i, $1 \leqq i \leqq n$ such that $x_i = -1$ and $y_i = 1$ and also $x_j = y_j$ for all j, such that $i < j \leqq n$. The following theorem gives the construction for COMPARISON.

THEOREM 2. *Let $m_k(X, Y) = P(x_n y_n, x_{n-1} y_{n-1}, \ldots, x_{k+1} y_{k+1})$ and let $m_n(X, Y) = q - n$, where $P(\cdot)$ is the delta polynomial described in Theorem 1 with $q \geqq n^4$ an odd prime power. Define*

$$\hat{C}(X, Y) = m_0(X, Y) + \sum_{i=1}^{n} (y_i - x_i) m_i(X, Y).$$

Then $C(X, Y) = \operatorname{sgn}(-\hat{C}(X, Y))$.

Proof. We consider the two cases ($X \geqq Y$ or $X < Y$) and prove that $C(X, Y) = \operatorname{sgn}(-\hat{C}(X, Y))$ in both cases.

First, assume that X is strictly greater than Y. Hence, there is an i such that $x_i = -1$ and $y_i = 1$ and also $x_j = y_j$ for all j, $i < j \leqq n$. Hence, $(y_i - x_i)m_i \geqq 2(q - n)$ and

$$\hat{C}(X, Y) \geqq 2(q - n) - 2n(n - 1)\sqrt{q} > 0.$$

If $X = Y$, then clearly $\hat{C}(X, Y) = q - n > 0$. Hence, if $X \geqq Y$, then $-1 = C(X, Y) = \operatorname{sgn}(-\hat{C}(X, Y))$.

Similarly, if $X < Y$, then $\hat{C}(X, Y) \leqq -2(q - n) + 2n(n - 1)\sqrt{q} < 0$. Hence, $C(X, Y) = \operatorname{sgn}(-\hat{C}(X, Y))$ in this case as well, completing the proof. □

Next, we consider the ADDITION function. To compute the bits of the sum of the two n-bit numbers X and Y as signs of sparse polynomials, it suffices to construct a sparse polynomial for each of the carry bits. This is because the ith bit in the result of the addition is $x_i y_i c_i$, where c_i is the corresponding carry bit. If we can compute c_i as a sign of a sparse polynomial, say $c_i = \operatorname{sgn}(p(X, Y))$, then we can also compute $x_i y_i c_i = \operatorname{sgn}(x_i y_i p(X, Y))$ as a sign function of a sparse polynomial. We will henceforth concentrate, without loss of generality, on proving that the carry to the last bit (i.e., c_n) can be computed as a sign of a sparse polynomial. We denote the carry function to the last bit as CAR (X, Y) and prove that it can be computed as a sign function of a sparse polynomial.

THEOREM 3. *Let* $l_k(X, Y) = P(-x_{n-1}y_{n-1}, -x_{n-2}y_{n-2}, \ldots, -x_{k+1}y_{k+1})$ *and let* $l_{n-1}(X, Y) = q - n$, *where* $P(\cdot)$ *is the delta polynomial described in Theorem* 1 *with* $q \geqq 4n^4$ *an odd prime power. Let* $f_1(w_1, w_2) = (1 - w_1 - w_2 + w_1 w_2)$. *Let*

$$\widehat{\mathrm{CAR}}\,(X, Y) = \sum_{i=1}^{n-1} f_1(x_i, y_i) l_i(X, Y).$$

Then CAR $(X, Y) = \operatorname{sgn}(2q - \widehat{\mathrm{CAR}}\,(X, Y))$.

Proof. Note that $f_1(-1, -1) = 4$ and $f_1(1, 1) = f_1(1, -1) = f_1(-1, 1) = 0$.

First, assume that there is carry to bit n in the addition of X and Y, namely, that CAR $(X, Y) = -1$. In such a case, we have carry generation and propagation. Namely, there is an $i < n$ such that $x_i = -1$ and $y_i = -1$ in which the carry is generated, and, in addition, $x_j \neq y_j$ for all j, $i < j < n$ (so that the carry will propagate). Note that the carry will also propagate in the case where $x_j = y_j = -1$. However, without loss of generality, we can consider the leftmost place i in which the carry was generated. Since $f_1(x_i, y_i)l_i \geqq 4(q - n)$, then, by the properties of the delta polynomials,

$$\widehat{\mathrm{CAR}}\,(X, Y) \geqq 4(q - n) - 4(n - 2)(n - 1)\sqrt{q} > 2q.$$

Hence, if there is carry, then CAR $(X, Y) = \operatorname{sgn}(2q - \widehat{\mathrm{CAR}}\,(X, Y))$.

Next, we consider the case in which there is no carry. The reason for not having a carry is that, for each index i, either there is no carry generation (and then $f_1(x_i, y_i) = 0$) or there is a carry generation but there is no carry propagation. In the latter case, $|l_i(X, Y)| \leqq (n - 1)\sqrt{q}$. Hence, for this case

$$\widehat{\mathrm{CAR}}\,(X, Y) \leqq 4(n - 1)^2\sqrt{q} < 2q.$$

Hence, if there is no carry, then CAR $(X, Y) = \operatorname{sgn}(2q - \widehat{\mathrm{CAR}}\,(X, Y))$, completing the proof. □

4. Concluding remarks and extensions. A family of vectors F in $\{-1, 1\}^n$ is a *linear subspace* if, for every $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in F, the vector $x*y = (x_1y_1, \ldots, x_ny_n)$ is also in F. (This is the usual definition of a subspace together with our mapping that replaces 0 and 1 by 1 and -1, respectively.) Similarly, A is an *affine subspace* if it is the set of all vectors of the form $x*y$ for some fixed vector x as y ranges over all vectors of a linear subspace. Generalizing the notion of a delta polynomial, we can construct, for every affine subspace, a sparse polynomial whose value on the members of the subspace is much larger than one whose value on vectors is outside the subspace. (The delta polynomials correspond to the case that the subspace contains only one point.) This enables us, among other things, to explicitly express every function that is the characteristic function of a union of polynomially many affine subspaces as a sign of a sparse polynomial. To construct the generalized delta polynomials, we first observe that it suffices to construct those for linear subspaces. For every linear subspace of codimension k in $\{-1, 1\}^n$, there are k monomials in $x_1, \ldots, x_n$ such that a vector $(x_1, \ldots, x_n)$ is in the subspace if and only if all these monomials evaluated in the coordinates of the above vector are 1. We can thus simply substitute these monomials in the delta polynomial of §2 and obtain the desired generalized sparse polynomial. We omit the details.

The delta polynomials can be used to construct a sparse polynomial for the MAXIMUM function, which gets as input n integers (n bits each) and outputs -1 if and only if the first integer is the maximum. The construction for MAXIMUM is a simple generalization of the one for the COMPARISON function.

By a nonconstructive argument, we can prove that there is a $(q - n)$-sparse polynomial $P(x_1, \ldots, x_n)$ satisfying somewhat stronger properties than those given by Theorem 1; namely,

(i) $P(1, 1, \ldots, 1) = q - n$ and

(ii) For every $(x_1, \ldots, x_n) \in \{-1, 1\}^n$, which is not $(1, 1, \ldots, 1)$, $|P(x_1, \ldots, x_n)| \leqq O(n^{1/2}q^{1/2})$.

It would be interesting to find an explicit construction of such polynomials. (This will supply, of course, smaller depth-2 MAJ-circuits for the functions considered in §3.)

REFERENCES

[1] M. Ajtai, J. Komlós, and E. Szemerédi, *Deterministic simulation in* LOGSPACE, in Proc. 19th Annual ACM STOC, ACM Press, New York, 1987, pp. 132–140.

[2] N. Alon, *Tools from higher algebra*, Handbook of Combinatorics, R. L. Graham, M. Grotschel, and L. Lovász, eds., North-Holland, Amsterdam, to appear.

[3] N. Alon, O. Goldreich, J. Hastad, and R. Peralta, *Simple constructions of almost k-wise independent random variables*, in Proc. 31st IEEE FOCS, St. Louis, MO, IEEE (1990), pp. 544–553; Random Structures Algorithms, 3 (1992), pp. 289–304.

[4] E. Bach, *Realistic analysis of some randomized algorithms*, in Proc. 19th Annual ACM STOC, ACM Press, New York, 1987, pp. 453–461.

[5] B. Bollobás, *Random Graphs*, Academic Press, London, 1985.

[6] J. Bruck, *Harmonic analysis of polynomial threshold functions*, SIAM J. Discrete Math., 3 (1990), pp. 168–177.

[7] J. Bruck and R. Smolensky, *Polynomial threshold functions, AC^0 functions and spectral norms*, SIAM J. Comput., 21 (1992), pp. 33–42.

[8] A. K. Chandra, L. Stockmeyer, and U. Vishkin, *Constant depth reducibility*, SIAM J. Comput., 13 (1984), pp. 423–439.

[9] R. L. Graham and J. H. Spencer, *A constructive solution to a tournament problem*, Canad. Math. Bull., 14 (1971), pp. 45–48.

[10] T. Hofmeister, T. Hohberg, and S. Kohling, *Some notes on threshold circuits and multiplication in depth* 4, Inform. Process. Lett., 39 (1991), pp. 219–225.

[11] A. Hajnal, W. Maass, P. Pudlak, M. Szegedy, and G. Turan, *Threshold circuits of bounded depth*, in Proc. 28th IEEE FOCS, 1987, pp. 99–110.

[12] J. Naor and M. Naor, *Small-bias probability spaces: Efficient constructions and applications*, in Proc. 22nd Annual ACM STOC, 1990, ACM Press, New York, pp. 213–223.

[13] R. Peralta, *On the Randomness Complexity of Algorithms*, CS Research Report TR 90-1, Univ. of Wisconsin, Milwaukee, WI.

[14] W. M. Schmidt, *Equations Over Finite Fields, An Elementary Approach*, Springer Lecture Notes in Mathematics, Vol. 536, Springer-Verlag, Berlin, 1976.

[15] K. Y. Siu and J. Bruck, *On the power of threshold circuits with small weights*, SIAM J. Discrete Math., 4 (1991), pp. 423–435.

[16] K. Y. Siu, J. Bruck, T. Kailath, and T. Hofmeister, *Depth Efficient Neural Networks for Division and Related Problems*, IBM Research Report, RJ 7946, 1991.

[17] I. Wegener, *The Complexity of Boolean Functions*, John Wiley, New York, 1987, p. 322.

[18] A. Weil, *Sur les courbes algébriques et les variéstés qui sèn déduisent*, Actualités Sci. Ind. No. 1041, 1948.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 9-15, February 1994

THE NUMBER OF DEGREE-RESTRICTED ROOTED MAPS ON THE SPHERE*

EDWARD A. BENDER† AND E. RODNEY CANFIELD‡

Abstract. Let D be a set of positive integers. Let $m(n)$ be the number of n edged rooted maps on the sphere all of whose vertex degrees (or, dually, face degrees) lie in D. Using Brown's technique, the generating function for $m(n)$ implicitly is obtained. It is used to prove that, when gcd (D) is even,

$$m(n) \sim C(D)n^{-5/2}\gamma(D)^n.$$

It also yields known formulas for various special D.

Key words. root edge, power series, multiset

AMS subject classification. 05C30

1. Introduction. Let D be a set of positive integers containing some element exceeding 2, let $M(x, y) = \sum_i M_i(x)y^i$ be the generating function by edges and root face degree for rooted maps on the sphere such that each nonroot face degree lies in D, and let $m(n)$ be the number of n edged rooted maps all of whose face degrees lie in D. Define the coefficient operator with respect to y by

$$[y^k]\left(\sum_{i \geqq 0} f_i(x)y^i\right) = f_k(x)$$

and define $[x^k]$ similarly. We will prove the following theorem.

THEOREM 1. *There exist unique power series $R_1(x)$ and $R_2(x)$ such that*

$$R_1 = \frac{x}{2}\sum_{i \in D}[y^{i-1}](R^{-1/2}) \tag{1.1}$$

and

$$R_2 = \frac{x}{2}\sum_{i \in D}[y^i](R^{-1/2}) + x - 3R_1^2, \tag{1.2}$$

where $R = 1 - 4R_1y - 4R_2y^2$. We have

$$\begin{aligned} m(n) &= \frac{1}{n+1}[x^n](M_2'(x)) \\ &= [x^n]\left(\frac{(R_2(x) + R_1(x)^2)(R_2(x) + 9R_1(x)^2)}{(n+1)x^2}\right). \end{aligned} \tag{1.3}$$

Although the sums in (1.1) and (1.2) appear quite formidable, they can be simplified in some interesting cases. Two particularly simple situations are those in which gcd (D) is even and those related to arithmetic progressions and finite sets. We have the following corollaries.

* Received by the editors February 12, 1990; accepted for publication (in revised form) October 28, 1992. This research was supported by the National Security Agency. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation hereon.

† Department of Mathematics, University of Californa, San Diego, La Jolla, California 92093.

‡ Department of Computer Science, University of Georgia, Athens, Athens, Georgia 30602.

COROLLARY 1.1. *If, when viewed as a multiset, D differs from the union of a finite number of arithmetic progressions by a finite multiset, then $M(x, y)$ is algebraic.*

COROLLARY 1.2 (see Tutte [6]). *The number of $2d$ regular, nd edge, rooted maps on the sphere is*

$$\frac{2(nd)!}{n!(nd-n+2)!}\binom{2d-1}{d}^n.$$

COROLLARY 1.3 (see Liu [4]). *The number of n edge, rooted, bipartite (or, dually, Eulerian) maps on the sphere is*

$$\frac{3}{2(n+1)(n+2)}\binom{2n}{n}2^n.$$

THEOREM 2. *If* $\gcd(D) = 2d$, *then*

$$m(n) \sim \frac{2d(\sigma\gamma)^{5/2}}{(\pi\lambda)^{1/2}} n^{-5/2}\gamma^n,$$

where n is a multiple of d, σ is the positive real root of

$$2 = \sum_{2i\in D}(i-1)\binom{2i}{i}\sigma^i, \quad \lambda = \sum_{2i\in D} i(i-1)\binom{2i}{i}\sigma^i, \quad \gamma = \frac{1}{2}\sum_{2i\in D} i\binom{2i}{i}\sigma^{i-1}.$$

We will use a Tutte-type decomposition to obtain a quadratic equation for $M(x, y)$ when D is finite. Such equations are usually solved by the quadratic method [3]. That approach does not seem to work here. We must look more closely at what Brown's result [2] says about the discriminant of the quadratic. Having established Theorem 1 for finite D, we then pass to the limit. When $\gcd(D)$ is even, $R_1 = 0$. This results in considerable simplication of the equations in Theorem 1, which easily leads to Corollaries 1.2 and 1.3. It also leads to equations that suggest that there may be an interesting bijection between various bipartite maps and pairs of some other objects. We have not been able to find the bijection. In §4 we use a result of Meir and Moon [5] to obtain Theorem 2 from Theorem 1.

At this time, we have no general asymptotic result when $\gcd(D)$ is odd. We suspect that a result of the form (1.6) will hold. This can be verified on a case-by-case basis with lengthy calculations. For example, with D the set of odd positive integers, we have used Maple to prove that

$$m(n) \sim Cn^{-5/2}x_0^{-n},$$

where

$$x_0 = 0.105191\cdots, \qquad C = 3\tau^{1/2}/4\pi^{1/2} = 0.71772\cdots,$$

and both x_0 and τ are algebraic of degree 6.

2. Proof of Theorem 1. For all of this section, except the last paragraph, we assume that D is a finite set with largest element t.

Note that (1.1) tells us that R_1 has no constant term. If we specify R_1 and R_2 through terms of degree k in x and substitute them in the right sides of (1.1) and (1.2), the left sides give us R_1 and R_2 through degree $k+1$. Thus the power series R_1 and R_2 are uniquely determined by (1.1) and (1.2) and have no constant terms.

Let $\mathbf{C}$ denote the complex numbers, let $\mathscr{R}[[x]]$ denote formal power series over the commutative ring $\mathscr{R}$, and let $\mathscr{R}[y]$ denote polynomials. Let

$$\theta_k(y) = \sum_{\substack{i \in D \\ i \geqq k}} y^{t-i}. \tag{2.1}$$

Note that $\theta_k(y) \equiv \theta_0(y) \pmod{y^{t-k+1}}$.

We use a standard construction [7] to obtain a functional equation for $M(x, y)$. Either a map consists of just one vertex, with generating function 1, or it has a root edge. The generating function for maps for which the removal of the root edge leaves two components is given by $xy^2M(x, y)^2$. There is one more case, namely, removing the root edge does not disconnect the map. Reversing these removals gives a recursive construction for the maps. If the root face has degree j and if we wish to add a new root that reverses the last case, we can create a nonroot face of degree k and a root face of degree $j + 2 - k$. This construction leads to

$$M(x, y) = 1 + xy^2M(x, y)^2 + x \sum_{j \geqq 0} M_j(x) \sum_{\substack{k \in D \\ k \leqq j+1}} y^{j+2-k}.$$

After some algebra,

$$\begin{aligned} M(x, y) = 1 + xy^2M(x, y)^2 + xy^{2-t}\theta_0(y)M(x, y) \\ - x \sum_{j=0}^{t-2} \theta_{j+2}(y)y^{j+2-t}M_j(x). \end{aligned} \tag{2.2}$$

It is important to note that the recursive nature of this construction guarantees that *there is a unique power series solution to* (2.2).

Regarding (2.2) as a quadratic in $M(x, y)$, we obtain

$$2xy^tM(x, y) = y^{t-2} - x\theta_0(y) \pm B(x, y)^{1/2}, \tag{2.3}$$

where

$$B(x, y) = (x\theta_0(y) - y^{t-2})^2 - 4xy^{2t-2} + 4x^2y^t \sum_{j=0}^{t-2} \theta_{j+2}(y)M_j(x)y^j. \tag{2.4}$$

We show that (2.3) is true algebraically modulo y^{2t-1} for any $M_j(x) \in \mathbf{C}[[x]]$, provided that $M_0(x) = 1$ and the proper sign is chosen for the square root. (By "algebraically," we mean that no map information is needed.) To see this, first rearrange (2.3) as

$$\pm B(x, y)^{1/2} \equiv x\theta_0(y) - y^{t-2} + 2xy^tM(x, y) \qquad \pmod{y^{2t-1}}$$

and note that the right side contains a nonzero term in y^0. Thus this congruence holds if and only if

$$\begin{aligned} B(x, y) &\equiv (x\theta_0(y) - y^{t-2} + 2xy^tM(x, y))^2 \pmod{y^{2t-1}} \\ &\equiv (x\theta_0(y) - y^{t-2})^2 + 4x^2y^t\theta_0(y)M(x, y) - 4xy^{2t-2}M_0(x) \pmod{y^{2t-1}}, \end{aligned}$$

which is easily verified using (2.4).

From the previous paragraph, it follows that *any* choice for $M_1(x), \ldots, M_{t-2}(x) \in \mathbf{C}[[x]]$ that leads to a power series for $B(x, y)^{1/2}$ will be a solution to (2.3) and hence the unique solution.

Since $B(x, y) \in \mathbf{C}[[x]][y]$ has degree $2t - 2$ in y, Brown's theorem [2] guarantees that $B(x, y) = Q(x, y)^2 R(x, y)$, for some $Q, R \in \mathbf{C}[[x]][y]$ with $R(x, 0) = 1$. We will show that one solution can be found with

$$(2.5)\quad Q(x, y) = x + \sum_{i=1}^{t-2} Q_i(x) y^i \quad \text{and} \quad R(x, y) = 1 - 4R_1(x)y - 4R_2(x)y^2.$$

To begin, Q^2R has the same degree as B with respect to y, namely, $2t - 2$. Since (2.3) is true algebraically modulo y^{2t-1}, it suffices to determine the t unknown functions R_1, $R_2, Q_1, \ldots, Q_{t-2}$ by looking at the coefficients of y, y^2 up to y^t in

$$(2.6)\qquad 2xy^t M(x, y) = y^{t-2} - x\theta_0(y) + Q(x, y)R(x, y)^{1/2}$$

and then showing that $M_1(x), \ldots, M_{t-2}(x)$ are, in fact, power series.

It follows from (2.6) that

$$(2.7)\qquad Q(x, y) \equiv R(x, y)^{-1/2}(x\theta_0(y) - y^{t-2} + 2xy^t) \qquad (\text{mod } y^{t+1}).$$

Since we defined Q to have degree $t - 2$ in y, reduction modulo y^{t-1} determines Q in terms of R_1 and R_2, while the coefficients of y^{t-1} and y^t in (2.7) give two polynomial equations in the two unknowns R_1 and R_2. These equations are (1.1) and (1.2). We have shown that $Q_i(x) \in \mathbf{C}[[x]]$ and $R_j \in \mathbf{C}[[x]]$ can be found. Since $R_1(x)$ and $R_2(x)$ have no constant terms and $Q(x, y)$ has degree $t - 2$ in y, it follows from (2.6) that $xM_j(x)$ has no constant term, and so $M_j(x)$ is a power series. This completes the proof that (2.5) and (2.7) determine the unique power series solution $M(x, y)$ to (2.3).

We now prove (1.3). Using G' to denote $\partial G/\partial x$, we have

$$(2.8)\qquad (QR^{1/2})' = \tfrac{1}{2} R^{-1/2}(2Q'R + QR'),$$

and from (2.7)

$$(2.9)\qquad (QR^{1/2})' \equiv \theta_0(y) + 2y^t \qquad (\text{mod } y^{t+1}).$$

Thus

$$(2.10)\qquad 2Q'R + QR' \equiv 2R^{1/2}(\theta_0(y) + 2y^t) \qquad (\text{mod } y^{t+1}).$$

Define the truncation operator with respect to y by

$$\mathcal{T}_k\Bigl(\sum_{i \geqq 0} f_i(x) y^i\Bigr) = \sum_{i=0}^{k} f_i(x) y^i.$$

Since the left-hand side of (2.10) equation has degree t in y,

$$2Q'R + QR' = \mathcal{T}_t(2R^{1/2}(\theta_0(y) + 2y^t)).$$

Thus

$$\begin{aligned} x(QR^{1/2})'R^{1/2} &= x\mathcal{T}_t(R^{1/2}(\theta_0(y) + 2y^t)) \\ &= \mathcal{T}_t(R^{1/2}(x\theta_0(y) - y^{t-2} + 2xy^t)) + \mathcal{T}_t(R^{1/2}y^{t-2}) \\ &= \mathcal{T}_t(RR^{-1/2}(x\theta_0(y) - y^{t-2} + 2xy^t)) + y^{t-2}\mathcal{T}_2(R^{1/2}) \\ &= RQ + y^{t-2}\mathcal{T}_2(R^{1/2}), \end{aligned}$$

by (2.7). Combining this with (2.3), we have

$$\begin{aligned} 2x(xM_j(x))' &= [y^{t+j}](x(QR^{1/2})') \\ &= [y^{t+j}](QR^{1/2}) + [y^{t+j}](R^{-1/2}y^{t-2}\mathscr{T}_2(R^{1/2})) \\ &= 2xM_j(x) + [y^{j+2}](R^{-1/2}\mathscr{T}_2(R^{1/2})). \end{aligned}$$

Rearranging the two ends of this equation gives us

$$M_j'(x) = \frac{1}{2x^2}[y^{j+2}](R^{-1/2}\mathscr{T}_2(R^{1/2})). \tag{2.11}$$

Set $m(0) = 1$ and let $M(x)$ be the generating function for $m(n)$. By removing the root edge from a map counted by $M_2(x)$, it is easily seen that

$$M(x) = M_2(x)/x. \tag{2.12}$$

This, combined with a bit of algebra on (2.11), gives us (1.3). The proof of Theorem 1 for finite D is complete.

Let D be arbitrary and define $D(t)$ to be those elements of D that do not exceed t. We may apply Theorem 1 to $D(t)$. If we replace t by $t' > t$, then it is simple to check that the terms of degree less than $t/2$ do not change in the formulas for R_1 and R_2. Thus we may simply let $t \to \infty$.

3. Simple applications of Theorem 1. The reader may carry out the calculations in Theorem 1 when D is the positive integers, thereby rederiving the generating function for all maps: In this case, (1.1) and (1.2) become

$$R_1 = \frac{x}{2}(1 - 4R_1 - 4R_2)^{-1/2},$$

$$R_2 = \frac{x}{2}((1 - 4R_1 - 4R_2)^{-1/2} - 1) + x - 3R_1^2.$$

Eliminating R_2 leads to the quartic equation

$$0 = (12R_1^2 - 2R_1 + x)(4R_1^2 - 2R_1 - x).$$

Since R_1 has no constant term and since the generating function for maps must have a positive real singularity, the correct solution is

$$R_1 = \frac{1 - \sqrt{1 - 12x}}{12}.$$

The equation for $M(x)$ follows easily from (1.3) and (2.12).

To prove Corollary 1.1, we observe that (1.1) and (1.2) can be summed for i in an arithmetic progression by using multisection of the series $R^{-1/2}$ with respect to y and then setting $y = 1$. Thus the equations of R_1 and R_2 are algebraic. Now use (2.6) and the value of Q from (2.7).

When gcd (D) is even, $R_1 = 0$. We can see this either by noting that $M(x, y)$ cannot have any odd degree terms in y or by noting that the assumption $R_1 = 0$ leads to a solution, and so must be the unique solution. Since $R_1 = 0$, we have

$$[y^{2i}](R^{-1/2}) = [y^{2i}]((1 - 4R_2y^2)^{-1/2}) = \binom{2i}{i}R_2^i. \tag{3.1}$$

Thus (1.1) becomes 0 = 0, (1.2) becomes

$$R_2 = \frac{x}{2} \sum_{2i \in D} \binom{2i}{i} R_2^i + x, \tag{3.2}$$

and (1.3) becomes

$$m(n) = \frac{1}{n+1}[x^n](M'_2(x)) = \frac{1}{n+1}[x^n](x^{-2}R_2(x)^2). \tag{3.3}$$

Since $\binom{2i}{i}$ is even, it follows from (3.1) that $R_2(x)$ has nonnegative integer coefficients. This, combined with $M'_2(x) = (R_2(x)/x)^2$ from (3.3), suggests that there is probably an interesting bijection between rooted maps with a distinguished edge and pairs of combinatorial objects when gcd (D) is even.

When $D = \{2d\}$, the sum in (3.2) has only one term, and Corollary 1.2 follows easily by Lagrange inversion.

To prove Corollary 1.3, note that (3.2) becomes

$$R_2 = \frac{x}{2}((1 - 4R_2)^{-1/2} - 1) + x.$$

After some algebra, we obtain

$$R_2(x) = \frac{4x + 1 - \sqrt{1 - 8x}}{8}, \tag{3.4}$$

where the minus sign was chosen on the square root because $R_2(0) = 0$. Thus

$$\frac{R_2(x)^2}{x^2} = \frac{8x^2 + 1}{32x^2} + \left(\frac{(1 - 8x)^{3/2}}{32x}\right)'.$$

By (2.12) and (3.3),

$$M(x) = M_2(x)/x = \frac{1}{4} - \frac{1}{32x^2} + \frac{(1 - 8x)^{3/2}}{32x^2} + \frac{3}{8x}.$$

Corollary 1.3 follows easily.

4. Proof of Theorem 2. Suppose that $2d =$ gcd (D). It follows from (3.2) that $R_2(x) = x + xw(x^d)$, where the power series $w(z)$ is determined by

$$w = F(z, w) = \frac{1}{2} \sum_{2id \in D} \binom{2id}{id} z^i (w + 1)^{id}. \tag{4.1}$$

Let (ρ, τ) be a positive real solution, if any, of the simultaneous equations (4.1) and $1 = F_w(z, w)$. With some algebra, $\rho(\tau + 1)^d$ is the positive real root of

$$2 = \sum_{2id \in D} (id - 1)\binom{2id}{id}(\rho(\tau + 1)^d)^i, \tag{4.2}$$

and $\tau > 0$ is then determined by (4.1).

We wish to apply Meir and Moon's theorem [5, Thm. 1] and thereby obtain asymptotics by using their correction to [1, Thm. 5]. If D is finite, $F(z, w)$ is analytic for all z and w and the conditions for [5, Thm. 1] are satisfied. Suppose that D is infinite. Then $F(z, w)$ is analytic for $4|z(w + 1)^d| < 1$, since $\binom{2n}{n} \sim 4^n/(\pi n)^{1/2}$ and (4.2) has a solution with $4\rho(\tau + 1) < 1$, since $(id - 1)\binom{2id}{id}$ is unbounded. Again [5, Thm. 1] applies.

Combining this with [1, (7.1)], we find that

$$[z^n](w+1) \sim ((\tau+1)/2\pi F_{ww}(\rho,\tau)d)^{1/2} n^{-3/2} \rho^{-n},$$

and w behaves like $\tau + 1 + C(1 - z/\rho)^{1/2}$ near $z = \rho$. Thus

$$[z^n]((w+1)^2) \sim 2(\tau+1)((\tau+1)/2\pi F_{ww}(\rho,\tau)d)^{1/2} n^{-3/2} \rho^{-n}. \tag{4.3}$$

In terms of the notation in Theorem 2, we find with some algebra that $\rho = \gamma^{-d}$, $\tau + 1 = \sigma\gamma$, and $F_{ww}(\rho, \tau) = \lambda/2(\sigma\gamma)^2$. Theorem 2 now follows easily from (3.3), $R_2/x = w + 1$, and (4.3).

REFERENCES

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

[2] W. G. BROWN, *On the existence of square roots in certain rings of power series*, Math. Ann., 158 (1965), pp. 82–89.

[3] ———, *An algebraic technique for solving certain problems in the theory of graphs*, in Theory of Graphs: Proceedings of the Colloquium Held at Tihany, Hungary, September 1966, P. Erdös and G. Katona, eds., Academic Press, New York, 1968, pp. 57–60.

[4] Y. LIU, *A note on the enumeration of bipartite planar maps*, Acta Math. Sci., to appear.

[5] A. MEIR AND J. W. MOON, *On an asymptotic method in combinatorics*, J. Combin. Theory Ser. A, 51 (1989), pp. 77–89.

[6] W. T. TUTTE, *A census of slicings*, Canad. J. Math., 14 (1962), pp. 708–722.

[7] ———, *On the enumeration of planar maps*, Bull. Amer. Math. Soc., 74 (1968), pp. 64–74.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 16–31, February 1994

003

THE GRAPH PARTITIONING POLYTOPE ON SERIES-PARALLEL AND 4-WHEEL FREE GRAPHS*

SUNIL CHOPRA†

Abstract. The graph partitioning polytope $P(G)$ is the convex hull of the incidence vectors of all partitions of a graph G. The authors show that $P(G)$ is completely defined by cycle inequalities if G is series-parallel and by cycle, 3-wheel, and repeated 2-sums of 3-wheel and cycle inequalities if G is a 4-wheel free graph.

Key words. graph partitioning, polytope, series-parallel, 4-wheel

AMS subject classifications. 05C40, 90C27

1. Introduction. Given a connected graph $G = (V, E)$, define $\pi = (V_i, i = 1, \dots, r)$ to be a *partition* if

$$|V_i| \geq 1, \qquad i \in \{1, \dots, r\},$$
$$|V_i \cap V_j| = 0 \quad \text{for } i \neq j \in \{1, \dots, r\}, \quad \text{and}$$
$$\cup_i V_i = V.$$

Define the edge set $E(\pi)$, where

$$E(\pi) = \{(u, v) \in E | \{u, v\} \not\subseteq V_i \quad \text{for } i \in \{1, \dots, r\}\}.$$

$E(\pi)$ is the set of edges with endpoints in two different subsets in the partition π. We also refer to $E(\pi)$ as the set of edges cut by the partition π. Given a partition π, define the incidence vector $x(\pi)$, where

$$x_e(\pi) = \begin{cases} 1 & \text{for } e \in E(\pi), \\ 0 & \text{otherwise.} \end{cases}$$

Let Π be the family of all partitions of the graph G. Consider edge weights $c_e \in R$ for all edges e in E. Define the weight of a partition $c(\pi)$, where $c(\pi) = \sum_{e \in E(\pi)} c_e$. The *unrestricted graph partitioning problem* (UGPP) is to find the minimum weight member of Π. Grötschel and Wakabayashi [8] studied this problem for the case when G is a complete graph. They refer to this case as the *clique partitioning problem*. They show that the problem is NP-hard.

Define the graph partitioning polytope $P(G)$, where

$$P(G) = \text{conv } \{x(\pi) | \pi \in \Pi\}.$$

Grötschel and Wakabayashi [8] gave several families of facet-defining inequalities for $P(G)$ for the case when G is a complete graph. However, a complete inequality description of $P(G)$ is unlikely to be found in general.

G is said to be a *simple graph* if it does not contain any parallel edges. The following result can be proved using standard techniques (see, for instance, [4]). It is stated here without proof.

PROPOSITION 1.1. *If G is a simple graph, the graph partitioning polytope $P(G)$ is full-dimensional.*

* Received by the editors May 20, 1991; accepted for publication (in revised form) February 9, 1993.

† Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208.

A graph with parallel edges is referred to as a *multigraph*. Let $\bar{G} = (V, \bar{E})$ be a multigraph with parallel edges e_1 and e_2. Define $G = (V, E)$, where $E = \bar{E} - \{e_2\}$. There is a very close relationship between the polyhedra $P(\bar{G})$ and $P(G)$. Assume that a complete inequality description of $P(G)$ is known and is given by

$$P(G) = \{a^i x \le a_i^0, x \in R_+^E, a^i \in R^E, a_i^0 \in R, i \in I\}.$$

Here I indexes the set of inequalities defining $P(G)$. For each i in I, define the vector $\bar{a}^i \in R^{\bar{E}}$, where

$$\bar{a}^i(e) = \begin{cases} a^i(e) & \text{for } e \in E, \\ 0 & \text{for } e = e_2. \end{cases}$$

PROPOSITION 1.2. *If $\bar{G}$ and G are as described above, a complete inequality description of $P(\bar{G})$ is given by*

$$\{\bar{a}^i x \le a_0^i, x(e_2) = x(e_1), x \in R_+^{\bar{E}}, \bar{a}^i \in R^{\bar{E}}, a_0^i \in R, i \in I\}.$$

Proof. Let $\bar{x}$ be any extreme point of the polyhedron

$$P_1(\bar{G}) = \{\bar{a}^i x \le a_0^i, x(e_2) = x(e_1), x \in R_+^{\bar{E}}, \bar{a}^i \in R^{\bar{E}}, a_0^i \in R, i \in I\}.$$

The proof follows from the observation that the projection of $P_1(\bar{G})$ on R^E is $P(G)$, and this projection is indeed a bijection between the two polytopes. □

Thus, when studying the polyhedron $P(G)$, we can restrict attention to simple graphs G. From Proposition 1.1, the associated polyhedron $P(G)$ is full-dimensional. *For the remainder of the paper, all graphs considered are simple graphs.*

We should also point out that, when studying $P(G)$, we restrict attention to 2-connected graphs. To the contrary, assume that $G = (V, E)$ is not 2-connected as shown in Fig. 1.1. Assume that $G_i = (V_i, E_i)$, $i = 1, 2$ and $V_1 \cap V_2 = \{\bar{v}\}$. Assume that a complete inequality description of the polytopes $P(G_i)$, $i = 1, 2$ is as given below:

$$P(G_1) = \{a^{1i} x^1 \le a_0^i \mid a^{1i} \in R^{E_1}, x^1 \in R_+^{E_1}, i \in I_1\},$$

$$P(G_2) = \{a^{2i} x^2 \le a_0^i \mid a^{2i} \in R^{E_2}, x^2 \in R_+^{E_2}, i \in I_2\}.$$

For each inequality in I_1, define $\bar{a}^{1i}$, where

$$\bar{a}_e^{1i} = \begin{cases} a_e^{1i} & \text{for } e \in E_1, \\ 0 & \text{for } e \in E_2. \end{cases}$$

For each inequality in I_2, define $\bar{a}^{2i}$, where

$$\bar{a}_e^{2i} = \begin{cases} a_e^{2i} & \text{for } e \in E_2, \\ 0 & \text{for } e \in E_1. \end{cases}$$

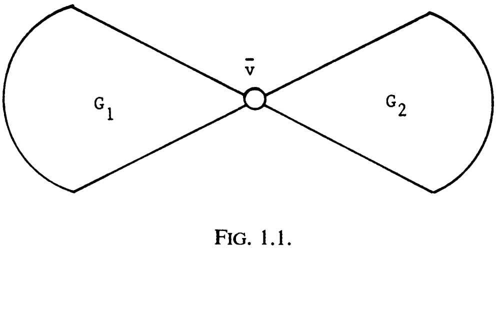


FIG. 1.1.

G is said to be the 1-*sum* of G_1 and G_2 ($G = G_1 +_1 G_2$). It is easy to verify the following result.

PROPOSITION 1.3. *If* $G_1 +_1 G_2$ *and* $P(G_1)$ *and* $P(G_2)$ *are as described above, then*

$$P(G) = \{x \in R_+^E \mid \bar{a}^{1i}x \le a_0^i \textit{ for } i \in I_1;\ \bar{a}^{2i}x \le a_0^i \textit{ for } i \in I_2\}.$$

For the remainder of the paper, we thus assume that *all graphs G are* 2-*connected*.

We assume familiarity with basic definitions in graph theory (see, for example, Bondy and Murty [3]). Consider two graphs $G_i = (V_i, E_i)$, $i = 1, 2$ with edges $e_i = (u_i, v_i) \in E_i$ for $i = 1, 2$. The $\bar{2}$-*sum* $G = G_1 \bar{+}_2 G_2$ is obtained by identifying the edges e_1 and e_2. The new node obtained on identifying $u_1(v_1)$ and $u_2(v_2)$ is called $u(v)$, and the new edge $\bar{e} = (u, v)$. All edges previously incident to u_1 or u_2 (v_1 or v_2) are now incident to $u(v)$.Thus, $G = (V, E)$, where $V = V_1 \cup V_2 \cup \{u, v\} - \{u_1, v_1, u_2, v_2\}$, $E = E_1 \cup E_2 \cup \{\bar{e}\} - \{e_1, e_2\}$. Define $G = G_1 +_2 G_2$ to be a 2-*sum* if G is formed as above, except that the edge $\bar{e}$ is deleted, i.e., $E = E_1 \cup E_2 - \{e_1, e_2\}$. The two operations are shown in Fig. 1.2.

A 2-*tree* can be defined recursively as follows. A triangle is a 2-tree. The $\bar{2}$-sum of a triangle and a 2-tree is a 2-tree. A graph is called *series-parallel* if it can be obtained from a forest by repeatedly adding an edge in parallel to an existing one or by replacing an edge by a path. From the above definitions, we obtain the following result (see Wald and Colburn [10]).

PROPOSITION 1.4. *Each simple, connected series-parallel graph is the subgraph of a* 2-*tree*.

A graph is called a 4-*wheel free graph* if it does not contain as a minor the graph W_4 shown in Fig. 1.3.

In this paper, we give a complete inequality description of $P(G)$ for the case when G is a series-parallel or 4-wheel free graph. The first result is given in § 2, and the second result in § 3.

Let $G = (V, E)$ be a connected graph. Consider a cycle $C = (V_c, E_c)$ in G. No partition π can cut exactly one edge from a cycle. For any edge $e^* \in E_c$, define the *cycle inequality*

$$(1.1) \qquad \sum_{e \in E_c - \{e^*\}} x_e - x_{e^*} \ge 0.$$

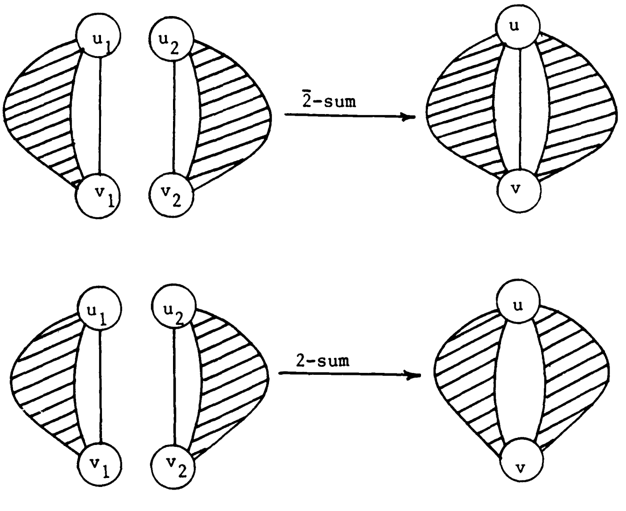


FIG. 1.2.

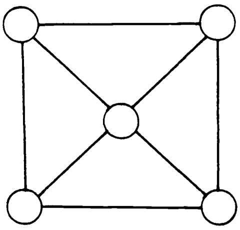


FIG. 1.3.

The cycle inequality is valid for $P(G)$. Grötschel and Wakabayashi [8] used these inequalities for triangles when formulating the clique partitioning problem. Chopra and Rao [4] showed the following result.

THEOREM 1.1. *The cycle inequality* (1.1) *is facet-defining for* $P(G)$ *if and only if* $|E_C| \geq 3$ *and* C *is a chordless cycle.*

Note that, for each cycle C, there are $|E_c|$ cycle inequalities. However, we write only one to represent all of them. Define the polyhedron $LP(G)$, where

$$LP(G) = \{x \in R^E \mid x \text{ satisfies } (1.1) \text{ for all cycles } C; \quad 0 \leq x_e \leq 1 \text{ for all } e \in E\}. \tag{1.2}$$

Chopra and Rao [4] showed the following result.

THEOREM 1.2. *It holds that* $P(G) = \operatorname{conv}\{x \in LP(G) \mid x \text{ integer}\}$.

2. $P(G)$ on series-parallel graphs. In general, $P(G) \subseteq LP(G)$; $P(G) \neq LP(G)$. The main result in this section is the following.

THEOREM 2.1. *If* G *is a connected series-parallel graph,* $P(G) = LP(G)$.

First, we give a general result on polytopes of graphs obtained by taking $\bar{2}$-sums of graphs whose polytopes are known. Given graphs $G_i = (V_i, E_i)$ with $e_i \in E_i$, $i = 1, 2$, identify the edges e_1 and e_2 into the edge $\bar{e}$ to get $G = G_1 \mp_2 G_2$, where $G = (V, E)$ and $E = E_1 \cup E_2 \cup \{\bar{e}\} - \{e_1, e_2\}$. Assume that complete inequality descriptions of $P(G_1)$ and $P(G_2)$ are known. For each inequality

$$a^1 x \leq a_0^1 \tag{2.1}$$

defining $P(G_1)$, construct the inequality

$$ax \leq a_0, \tag{2.2}$$

where $a_0 = a_0^1$ and

$$a_e = \begin{cases} a_e^1 & \text{for } e \in E_1 - \{e_1\}, \\ a_{e_1}^1 & \text{for } e = \bar{e}, \\ 0 & \text{otherwise.} \end{cases}$$

In a similar fashion, for each inequality

$$a^2 x \leq a_0^2 \tag{2.3}$$

defining $P(G_2)$, construct the inequality

$$ax \leq a_0, \tag{2.4}$$

where $a_0 = a_0^2$, and

$$a_e = \begin{cases} a_e^2 & \text{for } e \in E_2 - \{e_2\}, \\ a_{e_2}^2 & \text{for } e = \bar{e}, \\ 0 & \text{otherwise.} \end{cases}$$

Both inequalities (2.2) and (2.4) are valid for $P(G)$. Let $\bar{P}(G)$ be the polytope defined by all inequalities of the form (2.2) or (2.4).

PROPOSITION 2.1. *If* $G = G_1 \mp_2 G_2$ *and* $\bar{P}(G)$ *is as defined above, then* $P(G) = \bar{P}(G)$.

Proof. Inequalities (2.2) and (2.4) include all the chordless cycle inequalities arising from G. From Theorem 1.2, it follows that the integer points of $\bar{P}(G)$ are precisely the vertices of $P(G)$. Assume that $\bar{P}(G)$ has a fractional vertex $\bar{x}$. Define $\bar{x}_i$, $i = 1, 2$, where

$$\bar{x}_i(e) = \begin{cases} \bar{x}_e & \text{for } e \in E_i - \{e_i\}, \\ \bar{x}_{\bar{e}} & \text{for } e = e_i. \end{cases}$$

We show that $\bar{x}_i$ is a vertex of $P(G_i)$, $i = 1, 2$. Assume that $\bar{x}_1$ and $\bar{x}_2$ are not vertices. Then we have

$$\bar{x}_1 = \sum \alpha_i y^i; \qquad \bar{x}_2 = \sum \beta_i z^i, \tag{2.5}$$

where $\alpha_i, \beta_i \geq 0$, $\sum \alpha_i = \sum \beta_i = 1$, and y^i and z^i are vertices of $P(G_1)$ and $P(G_2)$, respectively. Define

$$S_1 = \{i \mid y^i(e_1) = 1\}; \qquad S_2 = \{i \mid z^i(e_2) = 1\},$$
$$\bar{S}_1 = \{i \mid y^i(e_1) = 0\}; \qquad \bar{S}_2 = \{i \mid z^i(e_2) = 0\}.$$

Vectors in S_i are incidence vectors of partitions that cut e_i, while vectors in $\bar{S}_i$ are incidence vectors of partitions that do not cut e_i, $i = 1, 2$. Since $\bar{x}_1(e_1) = \bar{x}_2(e_2) = \bar{x}(\bar{e})$, we have

$$\sum_{i \in S_1} \alpha_i = \sum_{i \in S_2} \beta_i = \bar{x}(\bar{e}). \tag{2.6}$$

Order the indices in S_1 as $\{i(1), i(2), \ldots, i(k)\}$ for $k = |S_1|$, and those in $\bar{S}_1$ as $\{i(k+1), \ldots, i(\bar{k})\}$ for $|\bar{S}_1| = \bar{k} - k$. Order the indices in S_2 as $\{l(1), \ldots, l(m)\}$, and those in $\bar{S}_2$ as $\{l(m+1), \ldots, l(\bar{m})\}$ for $|S_2| = m$ and $|\bar{S}_2| = \bar{m} - m$.

Given $y^{i(r)}$ for $r \leq k$ and $z^{l(s)}$ for $s \leq m$, define $x^{r,s}$, where

$$x_e^{r,s} = \begin{cases} y_e^{i(r)} & \text{for } e \in E_1 - \{e_1\}, \\ z_e^{l(s)} & \text{for } e \in E_2 - \{e_2\}, \\ 1 & \text{for } e = \bar{e}. \end{cases}$$

For $k + 1 \leq r \leq \bar{k}$ and $m + 1 \leq s \leq \bar{m}$, define $x^{r,s}$, where

$$x_e^{r,s} = \begin{cases} y_e^{i(r)} & \text{for } e \in E_1 - \{e_1\}, \\ z_e^{l(s)} & \text{for } e \in E_2 - \{e_2\}, \\ 0 & \text{for } e = \bar{e}. \end{cases}$$

For the next definition, assume that $\alpha_{i(0)} = \beta_{l(0)} = 0$. Let $\delta(r, s) = \sum_{t=0}^{r} \alpha_{i(t)} - \sum_{t=0}^{s} \beta_{l(t)}$ for $r \in \{0, 1, \dots, \bar{k}\}$, $s \in \{0, 1, \dots, \bar{m}\}$. Define $\gamma(r, s)$ for $r \in \{1, \dots, \bar{k}\}$, $s \in \{1, \dots, \bar{m}\}$, where

$$\gamma(r, s) = \begin{cases} \beta_{i(s)} & \text{if } \delta(r-1, s-1) \le 0,\ \delta(r, s-1) \ge 0,\ \delta(r, s) \ge 0, \\ \delta(r, s-1) & \text{if } \delta(r-1, s-1) \le 0,\ \delta(r, s-1) \ge 0,\ \delta(r, s) \le 0, \\ \alpha_{i(r)} & \text{if } \delta(r-1, s-1) \ge 0,\ \delta(r-1, s) \le 0,\ \delta(r, s) \le 0, \\ -\delta(r-1, s) & \text{if } \delta(r-1, s-1) \ge 0,\ \delta(r-1, s) \le 0,\ \delta(r, s) \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\delta(k, m) = 0$ using (2.6) and the definition of $\delta(r, s)$. Thus, $\sum_{s=1}^{m} \sum_{r=1}^{k} \gamma(r, s) = \bar{x}(\bar{e})$ and $\sum_{s=m+1}^{\bar{m}} \sum_{r=k+1}^{\bar{k}} \gamma(r, s) = 1 - \bar{x}(\bar{e})$. Note that each vector $x^{r,s}$ defined above is an element of $\bar{P}(G)$, $\gamma(r, s) \ge 0$, and $\sum_s \sum_r \gamma(r, s) = 1$. From (2.6) and the orderings we have chosen, it follows that $\gamma(r, s) = 0$ for $r \le k$, $s \ge m + 1$ or $r \ge k + 1$, $s \le m$. Define $\tilde{x}$, where

$$\tilde{x} = \sum_{s=1}^{m} \sum_{r=1}^{k} \gamma(r, s) x^{r,s} + \sum_{s=m+1}^{\bar{m}} \sum_{r=k+1}^{\bar{k}} \gamma(r, s) x^{r,s}.$$

From (2.5) and the construction of $\bar{x}$, $x^{r,s}$, and γ, we have $\tilde{x}(e) = \bar{x}_i(e)$ for $e \in E_i$, $i = 1, 2$. Thus, $\tilde{x} = \bar{x}$. This contradicts our assumption that $\bar{x}$ is a vertex of $\bar{P}(G)$. Thus, either $\bar{x}_1$ is a fractional vertex of $P(G_1)$ or $\bar{x}_2$ is a fractional vertex of $P(G_2)$, a contradiction to the definition of $P(G_i)$, $i = 1, 2$. The result thus follows. □

PROPOSITION 2.2. *If G is a 2-tree, $P(G) = LP(G)$.*

Proof. For a triangle, we can verify that $P(G) = LP(G)$. The result follows from the definition of a 2-tree and the repeated application of Proposition 2.1. □

From Proposition 1.4, each simple, connected series-parallel graph $G = (V, E)$ is the subgraph of a 2-tree $\bar{G} = (V, \bar{E})$. Thus, $LP(G)$ is the projection of $LP(\bar{G})$ onto the set $\{x \mid x_e = 0\ \forall e \in \bar{E} - E\}$. We analyze this projection using techniques discussed in Balas and Pulleyblank [1]. Define

$$Z = \{(y, x) \in R^{p+q} \mid Ay + Hx \le b,\ y \ge 0,\ x \in Q\}$$

and its projection

$$X = \{x \in R^q \mid \exists y \in R^p \text{ s.t. } (y, x) \in Z\}.$$

Here $Q \subseteq R^q$. Let W be the cone $W = \{t \in R^m \mid tA \ge 0,\ t \ge 0\}$ and extr (W) the set of extreme rays of W. Then the projection X can be obtained as follows:

$$X = \{x \in R^q \mid (tH)x \le tb\ \forall t \in W;\ x \in Q\}. \tag{2.7}$$

In (2.7) we can restrict attention to the extreme rays $t \in$ extr (W).

We can rewrite the cycle inequalities in the following form:

$$-\sum_{e \in E_c - \{e^*\}} x_e + x_{e^*} \le 0. \tag{2.8}$$

Besides nonnegativity, the other defining inequalities for $LP(\bar{G})$ are

$$x_e \le 1 \quad \forall e \in \bar{E}. \tag{2.9}$$

In our case, $Z = LP(\bar{G})$, with A, the submatrix of inequalities (2.8) and (2.9) corresponding to the edges $\bar{E} - E$, and H, the submatrix corresponding to the edges in E.

PROPOSITION 2.3. *Consider any connected graph* $G = (V, E)$ *that is a subgraph of* $\bar{G} = (\bar{V}, \bar{E})$, *where* $\bar{E} - E = \{\bar{e}\}$. *Let* proj $(LP(\bar{G}))$ *be the projection of* $LP(\bar{G})$ $(LP(\bar{G})$ *is as defined in* (1.2)) *onto* R^E. *Then*

$$LP(G) = \text{proj}\,(LP(\bar{G})).$$

Proof. In this case, the matrix A is the column vector corresponding to the edge $\bar{e}$. The entry in this vector is 0, 1, or -1 for the rows corresponding to inequalities (2.8), and 0 or 1 for the rows corresponding to inequalities (2.9). The only row of type (2.9) with an entry of 1 in this column corresponds to $x(\bar{e}) \le 1$.

Let m be the total number of inequalities. Define the cone W, where

$$W = \{t \in R^m \mid tA \ge 0,\, t \ge 0\}.$$

Let S_i, $i \in \{-1, 0, 1\}$ be the set of rows of A with entry i in the column corresponding to $\bar{e}$. The extreme rays t of W are of the following form:

$$t_l = \begin{cases} 1 & \text{for } l = j \in S_0 \cup S_1, \\ 0 & \text{otherwise;} \end{cases} \tag{2.10}$$

$$t_l = \begin{cases} 1 & \text{for } l = j_1 \in S_{-1}, \\ 1 & \text{for } l = j_2 \in S_1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

Extreme rays of type (2.10) are unit vectors and do not alter the inequality on projection. Thus, the cycle inequality (2.8) is defining for proj $(LP(\bar{G}))$ if $\bar{e} \notin E_C$. Similarly, the upperbound inequalities (2.9) are defining for proj $(LP(\bar{G}))$ for $e \in E$.

Extreme rays of type (2.11) add two inequalities, one of which has a coefficient of $+1$ in $\bar{e}$, and the other has a coefficient of -1 in $\bar{e}$. If one of the inequalities is $x(\bar{e}) \le 1$ and the other a cycle inequality (2.8), the resulting inequality is a sum of $x_{e^*} \le 1$ and $-x_e \le 0$ for $e \in E_C - \{e^*, \bar{e}\}$ and is thus not facet-defining for proj $(LP(\bar{G}))$. If both inequalities are cycle inequalities given by cycles $C(1) = (V_{C(1)}, E_{C(1)})$ and $C(2) = (V_{C(2)}, E_{C(2)})$, where $\bar{e} \in E_{C(1)} \cap E_{C(2)}$, they have the following form:

$$-\sum_{e \in E_{c(1)} - \{\bar{e}\}} x_e + x_{\bar{e}} \le 0, \tag{2.12}$$

$$-\sum_{e \in E_{c(2)} - \{e^*\}} x_e + x_{e^*} \le 0 \quad \text{for } e^* \ne \bar{e}. \tag{2.13}$$

If $e^* \in E_{C(1)}$, the resulting inequality is the sum of the inequalities $-x_e \le 0$ for $e \in E_{C(1)} \cup E_{C(2)} - \{\bar{e}, e^*\}$ and can thus be dropped from the defining set. Thus, $e^* \notin E_{C(1)}$. First, consider the case where $\{E_{C(1)} \cap E_{C(2)}\} - \{\bar{e}\} = EI$ and $|EI| \ge 1$. The inequality defining proj $(LP(\bar{G}))$ is obtained by adding (2.12) and (2.13) and eliminating the column corresponding to $\bar{e}$. The resulting inequality can be written as a sum of the inequality

$$-\sum_{E_{C(1)} \cup E_{C(2)} - \{\bar{e}, e^*\}} x_e + x_{e^*} \le 0$$

and the inequalities $-x_e \le 0$ for $e \in EI$, each of which is valid for $LP(G)$. Thus, this inequality can be dropped from the defining set as $E_{C(1)} \cup E_{C(2)} - \{\bar{e}\}$ contains a cycle going through e^*.

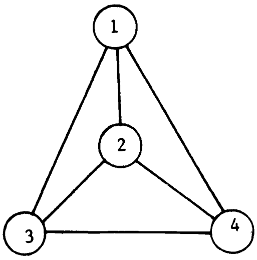


FIG. 2.1.

Now consider the case where $|EI| = 0$. Define $E_{\bar{C}} = E_{C(1)} \cup E_{C(2)} - \{\bar{e}\}$. The resulting inequality on projection is given by

$$(2.14) \qquad -\sum_{e \in E_{\bar{C}} - \{e^*\}} x_e + x_{e^*} \leq 0.$$

In this case, inequality (2.14) is a cycle inequality in G.

We have shown that proj $(LP(\bar{G}))$ is defined by inequalities of the form (2.8) if $\bar{e} \notin E_C$, (2.9), and (2.14), which are either defining inequalities for $LP(G)$ or can be written as a sum of such defining inequalities. Thus

$$\text{proj } (LP(\bar{G})) \subseteq LP(G).$$

Since each inequality defining $LP(G)$ is either defining for $LP(\bar{G})$ or is obtained on projection, it follows that $LP(G) \subseteq$ proj $(LP(\bar{G}))$. The result thus follows. □

Proof of Theorem 2.1. Since the case where parallel edges are present can be handled by Proposition 1.2, we consider $G = (V, E)$ to be a simple connected series-parallel graph. By Proposition 1.3, G is the subgraph of a 2-tree $\bar{G} = (\bar{V}, \bar{E})$. Repeated applications of Proposition 2.3 on the edges in $\bar{E} - E$ implies that $LP(G) =$ proj $(LP(\bar{G}))$. Proposition 2.2 implies that $P(\bar{G}) = LP(\bar{G})$. Also note that $P(G) =$ proj $(P(\bar{G}))$. Thus, we have $P(G) = LP(G)$. □

Remark 2.1. Theorem 2.1 is in some sense the best possible. Let G be the complete graph on four nodes. We give a fractional vertex of $LP(G)$. Every graph that is not series-parallel contains the complete graph on four nodes as a minor (see [10]). Furthermore, if $\bar{G}$ contains G as a minor, a vertex x of $LP(G)$ can be lifted to a vertex $\bar{x}$ of $LP(\bar{G})$ such that the restriction of $\bar{x}$ to G is x. It thus follows that, if $\bar{G}$ is not series-parallel, $LP(\bar{G})$ contains a fractional vertex. G is as shown in Fig. 2.1. A fractional vertex of $LP(G)$ is given by

$$x_{12} = x_{32} = x_{42} = \tfrac{1}{2}; \qquad x_{13} = x_{14} = x_{34} = 1.$$

Note that the example in Fig. 2.1 is a 4-wheel free graph. This case is treated in § 3.

Remark 2.2. UGPP can be solved in polynomial time on series-parallel graphs using the results of Bern, Lawler, and Wong [2]. Using the results of Grötschel, Lovasz, and Schrijver [7], Theorem 2.1 provides a polyhedral proof of this statement since cycle inequalities can be identified in polynomial time (see Deza, Grötschel, and Laurent [5]).

3. *P(G)* on 4-wheel free graphs. In this section, we give a complete inequality description of $P(G)$ for the case when G is a simple 4-wheel free graph. The instance with parallel edges can be resolved using Proposition 1.2 and the instance that is not 2-connected using Proposition 1.3. For the remainder of the section, we restrict our attention to 2-connected simple graphs G. The polytope $P(G)$ is thus full-dimensional by Proposition 1.1. From results of Seymour [9], it follows that a 2-connected simple, 4-wheel free graph $G = (V, E)$ can be obtained by a sequence of 2-sums or $\bar{2}$-sums of 2-connected simple, series-parallel graphs and copies of K_4, the complete graph on four nodes. Replacing each 2-sum in the above sequence by a $\bar{2}$-sum gives a graph $\bar{G} = (V, \bar{E})$. Clearly, $\bar{G}$ is also a 4-wheel free graph, and G is a subgraph of $\bar{G}$.

We first give a complete inequality description of $P(\bar{G})$. Let $K_4 = (V, E)$ be a complete graph on four nodes, where $V = \{1, 2, 3, 4\}$ and $E = \{12, 13, 14, 23, 24, 34\}$. For each node $i \in V$, define the *3-wheel inequality*

$$\sum_{j \neq l \in V - \{i\}} x_{jl} - \sum_{j \in V - \{i\}} x_{ij} \leq 1. \tag{3.1}$$

Chopra and Rao [4] showed these inequalities to be facet-defining for $P(K_4)$. The following result is given by Deza, Grötschel, and Laurent [6].

PROPOSITION 3.1. *The polyhedron $P(K_4)$ is completely defined by* (a) $0 \leq x_e \leq 1$ *for $e \in E$,* (b) *cycle inequalities* (1.1), *and* (c) *3-wheel inequalities* (3.1).

This allows us to show the following result.

PROPOSITION 3.2. *Let $\bar{G} = (V, \bar{E})$ be any graph obtained by repeated $\bar{2}$-sums of series-parallel graphs and copies of K_4. $P(\bar{G})$ is completely defined by* (a) $0 \leq x_e \leq 1$ *for $e \in E$,* (b) *cycle inequalities* (1.1), *and* (c) *3-wheel inequalities* (3.1).

Proof. The result follows by applying Proposition 2.1, Theorem 2.1, and Proposition 3.1. □

Next, we consider the operation of composing facets using 2-sums. Given $G_i = (V_i, E_i)$ with $e_i \in E_i$, $i = 1, 2$, let $G = G_1 +_2 G_2$ be the 2-sum obtained by identifying the edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ and deleting the resulting edge. Note that all graphs considered are simple and 2-connected. Consider inequalities

$$a^1 x \leq a_0^1, \tag{3.2}$$

$$a^2 x \leq a_0^2, \tag{3.3}$$

which are facet-defining for $P(G_1)$ and $P(G_2)$, respectively. Assume that neither (3.2) nor (3.3) is a bound inequality $-x_e \leq 0$ or $x_e \leq 1$. Furthermore, assume that $a^1(e_1) = -a^2(e_2) \neq 0$. Without loss of generality, assume that $a^1(e_1) > 0$. Define the inequality

$$ax \leq a_0, \tag{3.4}$$

where $a_0 = a_0^1 + a_0^2$ and

$$a_e = \begin{cases} a_e^1 & \text{for } e \in E_1 - \{e_1\}, \\ a_e^2 & \text{for } e \in E_2 - \{e_2\}. \end{cases}$$

PROPOSITION 3.3. *Inequality* (3.4) *is facet-defining for $P(G)$.*

Proof. Let π be any partition of V and let π_1 and π_2 be restrictions of π to V_1 and V_2, respectively. Either both edges e_1 and e_2 are cut by π_1 and π_2, respectively, or neither of the two edges are cut. Since $x(\pi_1)$ and $x(\pi_2)$ satisfy (3.2) and (3.3), respectively, and $a^1(e_1) = -a^2(e_2)$, it follows that $x(\pi)$ satisfies 3.4. Thus, inequality (3.4) is valid for $P(G)$.

Let x^1 and x^2 be equality solutions for (3.2) and (3.3), respectively, with $x^1(e_1) = x^2(e_2)$. Define the vector x, where

$$x_e = \begin{cases} x_e^1 & \text{for } e \in E_1 - \{e_1\}, \\ x_e^2 & \text{for } e \in E_2 - \{e_2\}. \end{cases}$$

$x \in P(G)$ is an equality solution to (3.4). Given the set of equality solutions to (3.2) and (3.3), construct all possible equality solutions to (3.4), as described above. If inequalities (3.2) and (3.3) are facet-defining, this gives a sufficient number of affinely independent equality solutions to (3.4) to show that it is facet-defining for $P(G)$. □

On forming a 2-sum of two K_4's, the 3-wheel inequalities can be composed with other 3-wheel inequalities. As an example, consider G, the 2-sum of two K_4's, shown in Fig. 3.1. Define $S = \{1, 2, 4, 8, 10\}$ and $T = \{3, 5, 6, 7, 9\}$. The inequality

$$\sum_{i \in S} x(e_i) - \sum_{i \in T} x(e_i) \leq 2 \tag{3.5}$$

is facet-defining for $P(G)$ using Proposition 3.3 since it is a 2-sum composition of two 3-wheel inequalities. Similarly, the inequality

$$x(e_4) + x(e_8) + x(e_{10}) - x(e_5) - x(e_6) - x(e_7) - x(e_9) \leq 1 \tag{3.6}$$

is obtained as a composition of a 3-wheel and a cycle inequality.

We can use these compositions to define a class WI of inequalities. The class WI is defined recursively as follows. 3-wheel inequalities are in WI; the 2-sum of two inequalities in WI or an inequality in WI, and a cycle inequality is also in WI.

The main result of this section is stated as Theorem 3.1.

THEOREM 3.1. *If* $G = (V, E)$ *is a simple, connected* 4-*wheel free graph*, $P(G)$ *is completely defined by* (a) $0 \leq x_e \leq 1$ *for* $e \in E$, (b) *cycle inequalities* (1.1), *and* (c) *inequalities in the class* WI.

Before proving Theorem 3.1, we need a few other results. Note once again that all graphs considered are simple. The polytope $P(G)$ is thus full-dimensional in each case. Consider the following condition:

(C1) Given the inequality $ax \leq a_0$ and an edge $\bar{e}$ with $a_{\bar{e}} < 0$, there exists a partition $\pi = (V_1, V_2)$ such that $\bar{e} \in E(\pi)$, $a_e \geq 0$ for $e \in E(\pi) - \{\bar{e}\}$ and $a_e > 0$ for at least one edge in $E(\pi) - \{\bar{e}\}$.

Consider the 2-sum operation described earlier with inequalities (3.2)–(3.4).

PROPOSITION 3.4. *If both inequalities* (3.2) *and* (3.3) *satisfy* (C1), *so does inequality* (3.4).

Proof. Consider any edge $e \in E_1 - \{e_1\}$ with $a_e^1 < 0$. By assumption, there exists a partition (V_1^1, V_1^2) of V_1 that satisfies (C1) for inequality (3.2) and e. If $\{u^1, v^1\} \subseteq V_1^1$, then $(V_1^1 \cup V_2 \cup \{u, v\} - \{u_1, v_1, u_2, v_2\}, V_1^2)$ is a partition that satisfies (C1) for (3.4) and e. If $\{u^1, v^1\} \subseteq V_1^2$, then $(V_1^1, V_1^2 \cup V_2 \cup \{u, v\} - \{u_1, v_1, u_2, v_2\})$ is a partition that satisfies (C1) for (3.4) and e. If $u_1 \in V_1^1$ and $u_2 \in V_1^2$, consider the partition (V_2^1, V_2^2) that satisfies (C1) for (3.3) and edge e_2 (earlier, we assumed, without loss of generality, that $a^2(e_2) < 0$). Assume that $u_2 \in V_2^1$ and $v_2 \in V_2^2$. The partition

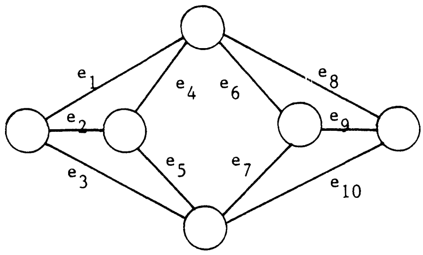


FIG. 3.1.

$(V_1^1 \cup V_2^1 \cup \{u\} - \{u_1, u_2\}, V_2^1 \cup V_2^2 \cup \{v\} - \{v_1, v_2\})$ satisfies (C1) for (3.4) and e. Since e is an arbitrarily chosen edge, the result follows. □

Note that both cycle inequalities and 3-wheel inequalities satisfy condition (C1). From Proposition 3.4 and the definition of the class of inequalities WI, we have the following result.

PROPOSITION 3.5. *All inequalities in the class* WI *satisfy condition* (C1).

Let $ax \le a_0$ be any inequality from the class WI for $P(G)$, where $G = (V, E)$. Assume that $a_{\bar{e}} = -1$. Consider the inequality

$$\bar{a}x \le \bar{a}_0, \tag{3.7}$$

where $\bar{a}_e = a_e$ for $e \in E - \{\bar{e}\}$ and $\bar{a}_0 = a_0 + 1$. Inequality (3.7) is valid for $P(\bar{G})$, where $\bar{G} = (V, E - \{\bar{e}\})$.

PROPOSITION 3.6. *Inequality* (3.7) *cannot be facet-defining for* $P(\bar{G})$.

Proof. From Proposition 3.5, there exists a partition $\pi = (V_1, V_2)$ of V that satisfies condition (C1) for $ax \le a_0$ and edge $\bar{e}$. Let $\bar{E}(\pi)$ be the edges cut by π in the graph $\bar{G}$. Define the edge set $E^+ = \{e \in \bar{E}_\pi | a_e > 0\}$. Note that $|E^+| \ge 1$, $E^+ \subseteq \bar{E}(\pi)$, and $a_e \ge 0$ for all $e \in \bar{E}(\pi)$. Consider any partition $\bar{\pi}$ whose incidence vector $x(\bar{\pi})$ satisfies (3.7) with equality. Let $\bar{E}(\bar{\pi})$ be the edges of $\bar{G}$ cut by $\bar{\pi}$. We show that $E^+ \subseteq \bar{E}(\bar{\pi})$. To the contrary, assume that $\tilde{e} \in E^+ - \bar{E}(\bar{\pi})$. Consider the partition $\tilde{\pi}$ obtained as the common refinement of π and $\bar{\pi}$. Let $\bar{E}(\tilde{\pi})$ be the edges in $\bar{G}$ cut by $\tilde{\pi}$. Clearly, $\bar{E}(\tilde{\pi}) = \bar{E}(\pi) \cup \bar{E}(\bar{\pi})$. Let $\tilde{x}$ be the incidence vector of $\bar{E}(\tilde{\pi})$. Since $\tilde{e} \in \bar{E}(\tilde{\pi})$, $\tilde{x}(\tilde{e}) = 1$. Also, $\bar{a}(\tilde{e}) > 0$ since $\tilde{e} \in E^+$. Thus,

$$\sum_{e \in \bar{E}} \bar{a}_e \tilde{x}_e > \sum_{e \in \bar{E}} \bar{a}_e \bar{x}_e = \bar{a}_0,$$

contradicting the validity of (3.7). Thus, each equality solution to (3.7) must be a partition that cuts all edges in E^+; i.e., every equality solution to (3.7) also satisfies the equation

$$\sum_{e \in E^+} x_e = |E^+|.$$

Thus, inequality (3.7) cannot be facet-defining for $P(\bar{G})$. □

Let $G = (V, E)$ be a 2-sum of $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ are identified with resulting nodes u and v. Consider a valid inequality for $P(G)$

$$ax \le a_0 \tag{3.8}$$

such that each partition satisfying (3.8) with equality has both nodes u and v in the same subset. Furthermore, assume that there are edges $\bar{e}_1 \in E_1 - \{e_1\}$ and $\bar{e}_2 \in E_2 - \{e_2\}$ such that $a(\bar{e}_1) \ne 0$ and $a(\bar{e}_2) \ne 0$.

PROPOSITION 3.7. *Inequality* (3.8) *cannot be facet-defining for* $P(G)$.

Proof. Let a^i be the restriction of a onto $E_i - \{e_i\}$, $i = 1, 2$. Define

$$a_0^i = \max\{a^i x(\pi) | \pi \text{ is a partition of } V_i\}, \qquad i = 1, 2.$$

Since inequality (3.8) is valid for $P(G)$ and each partition satisfying (3.8) with equality has both nodes u and v in the same subset, at least one of a_0^1 and a_0^2 is defined by a partition that has u_1 and v_1 or u_2 and v_2 in the same subset. Without loss of generality, assume this to be a_0^1. Define the inequality

$$bx \le a_0^1, \tag{3.9}$$

where

$$b_e = \begin{cases} a_e & \text{for } e \in E_1 - \{e_1\}, \\ 0 & \text{otherwise.} \end{cases}$$

Inequality (3.9) is valid for $P(G)$, and each equality solution to (3.8) also satisfies (3.9) with equality. Thus, inequality (3.8) is not facet-defining, since $P(G)$ is full-dimensional. The result thus follows. □

Consider a simple graph $G = (V, E)$, where G is either K_4 or series-parallel. Consider two inequalities $a^i x \leq a_0^i$, $i = 1, 2$ that are facet-defining for $P(G)$. Furthermore, assume that there are two edges $e_1, e_2 \in E$ such that $a^i(e_j) \neq 0$, $i, j = 1, 2$, $a^1(e_1) = -a^2(e_1) = 1$.

PROPOSITION 3.8. *There exists an edge $e \in E$, such that no partition π satisfying $a^i x(\pi) = a_0^i$, $i = 1, 2$ cuts e.*

Proof. The only instances that can arise on a K_4 are if one is a 3-wheel inequality and the other a 3-wheel or cycle inequality. For a series-parallel graph, both inequalities are cycle inequalities. Each case is resolved in Fig. 3.2. The coefficient on an edge is written next to the edge, and the edge e is marked. Note that $a_e^i \neq 0$ for at least one of $i = 1, 2$. □

Consider an inequality $ax \leq a_0$ that is facet-defining for $P(G)$, where $G = (V, E)$. Define $\bar{E} = \{e \in E \mid a_e \neq 0\}$. Let $\bar{G} = (\bar{V}, \bar{E})$ be the graph induced by edges in $\bar{E}$. $\bar{G}$ is called the *support graph* of the inequality $ax \leq a_0$.

Consider the inequalities

$$a^1 x \leq a_0^1, \tag{3.10}$$

$$a^2 x \leq a_0^2, \tag{3.11}$$

which are facet-defining for $P(G)$, where G is a 4-wheel free graph. Assume that inequality (3.10) belongs to the class WI and that (3.11) either belongs to WI or is a cycle inequality. Note that the support graphs of both (3.10) and (3.11) are 2-connected.

Assume that e_1 and e_2 are two edges such that $a^i(e_j) \neq 0$, $i, j = 1, 2$, and $a^1(e_1) = -a^2(e_1) = 1$. Since G is a 4-wheel free graph, we can write G as a 2-sum (or $\bar{2}$-sum) of graphs $G_i = (V_i, E_i)$, $i = 1, \ldots, r$, i.e., $G = G_1 +_2 \cdots +_2 G_r$. Each of the graphs G_i is either a K_4 or series-parallel. Each of the facets (3.10) and (3.11) is a 2-sum of 3-wheel and cycle inequalities defined on some of the G_i, $i = 1, \ldots, r$; i.e., there exists $S_j \subseteq \{1, \cdots r\}$, $j = 1, 2$ and inequalities

$$a^{ij} x \leq a_0^{ij}, \qquad i = 1, 2, \quad j \in S_i \tag{3.12}$$

such that (3.10) is a 2-sum of the inequalities for $i = 1$ and (3.11) is a 2-sum of the inequalities for $i = 2$. Each of the inequalities in (3.12) is either a 3-wheel or cycle inequality that is facet-defining for $P(G_i)$.

Assume that e_1 (described earlier) is an edge of G_k, i.e., $e_1 \in E_k$. Clearly, $k \in S_1 \cap S_2$. Since $e_1 \in E$, it is not deleted as the result of a 2-sum. Thus, $e \notin E_j$ for $j \neq k$. Inequality (3.12) gives the corresponding inequalities

$$a^{ik} x \leq a_0^{ik}, \qquad i = 1, 2.$$

Note that $a^{1k}(e_1) = -a^{2k}(e_1) = 1$.

PROPOSITION 3.9. *There exists an edge $e_3 \in E_k - \{e_1\}$ such that $a^{ik}(e_3) \neq 0$, $i = 1, 2$.*

Proof. The support graphs (3.10) and (3.11) are obtained as 2-sums of the support graphs $G_{ij} = (V_{ij}, E_{ij})$, $j \in S_i$, $i = 1, 2$ of inequalities in (3.12). Each of these support

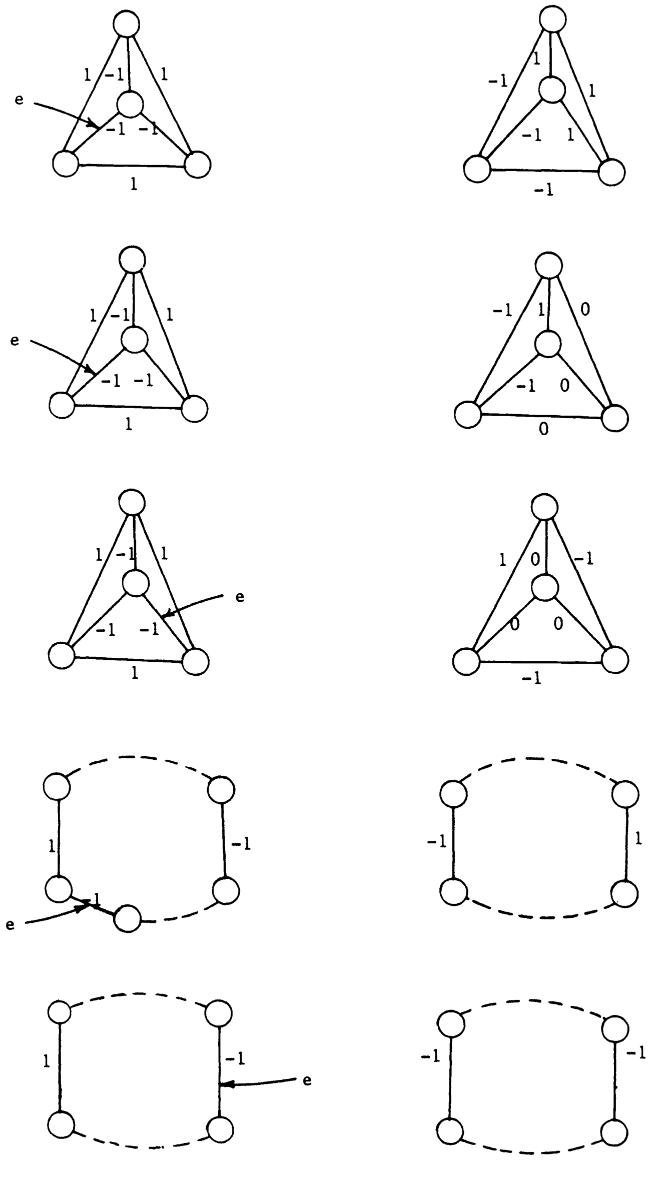


FIG. 3.2.

graphs is either a K_4 or a cycle. Assume that $E_{1k} \cap E_{2k} = \{e_1\}$. Since 2-sums are the only operations allowed and $e_1 \notin E_j$ for $j \neq k$, this implies that e_1 is the only edge shared by the support graphs of (3.10) and (3.11). This contradicts our assumption that they share e_2 as well. Thus, there exists an edge $e_3 \in \{E_{1k} \cap E_{2k}\} - \{e_1\}$. □

Given inequalities (3.10) and (3.11) described earlier, define the inequality

$$(3.13) \qquad ax \leq a_0,$$

where $a_0 = a_0^1 + a_0^2$ and $a_e = a_e^1 + a_e^2$ for $e \in E - \{e_1\}$. If $a_e = 0$ for all $e \in \bar{E}$, inequality (3.13) is not facet-defining for $P(G)$. Thus, we can assume that there exists some edge $\bar{e} \in E - \{e_1\}$ with $a(\bar{e}) \neq 0$. Define the graph $\bar{G} = (V, \bar{E})$, where $\bar{E} = E - \{e_1\}$. Inequality (3.13) is clearly valid for $P(\bar{G})$.

PROPOSITION 3.10. *Inequality* (3.13) *is not facet-defining for* $P(\bar{G})$.

Proof. The proof is by induction on the graph $\bar{G}$. If $\bar{G}$ is a K_4 or series-parallel, then Proposition 3.10 follows from Proposition 3.8. Thus $\bar{G}$ is obtained as a sequence of 2-sums (or $\bar{2}$-sums) of graphs G_i, $i = 1, \ldots, r$, discussed earlier. Each graph G_i in this sequence is either a K_4 or series-parallel. Thus, inequalities (3.10) and (3.11) are obtained as 2-sums of inequalities in (3.12). From Proposition 3.9, we know that there exists an index k such that $\{e_1, e_3\} \subseteq E_k$, $a^{ik}(e_1)$, $a^{ik}(e_3) \neq 0$, $a^{ik}(e_1) = -a^{2k}(e_1) = 1$, $i = 1, 2$. The graph $G_k = (V_k, E_k)$ is either a K_4 or series-parallel. From Proposition 3.8, there exists an edge $e_4 \in E_k$ such that no partition π_k of V_k satisfying $a^{ik}x(\pi_k) = a_0^{ik}$, $i = 1, 2$ cuts e_4. Assume that $e_4 = (u, v)$. Each partition π of V satisfying both (3.10) and (3.11) with equality is an extension of a partition π_k of V_k that satisfies $a^{ik}x(\pi_k) = a_0^{ik}$, $i = 1, 2$. Thus, each partition π of V satisfying both (3.10) and (3.11) with equality has u and v in the same subset. Note that, if a partition π satisfies (3.13) with equality, it satisfies both (3.10) and (3.11) with equality. If the edge e_4 is an edge in $\bar{E}$, then inequality (3.13) cannot be facet-defining, since there is no equality solution that cuts e_4; i.e., every equality solution satisfies $x(e_4) = 0$.

Now consider the case where e_4 is deleted during one of the 2-sum operations when forming G. Thus, G can be written as a 2-sum of $\tilde{G}_1 = (\tilde{V}_1, \tilde{E}_1)$ and $\tilde{G}_2 = (\tilde{V}_2, \tilde{E}_2)$, where e_4 is the identified edge. As mentioned earlier, there exists an edge $\bar{e} \in \bar{E}$ with $a(\bar{e}) \neq 0$. Without loss of generality, assume that $\bar{e} \in \tilde{E}_1 - \{e_4\}$. $\tilde{G}_2$ itself is a sequence of 2-sums (or $\bar{2}$-sums) of K_4's and series-parallel graphs. Let $\tilde{S}_2 \subseteq \{1, \ldots, r\}$ be the index set of the graphs used to form $\tilde{G}_2$. One of the components in $\tilde{S}_2$ must contain the edge e_4. Without loss of generality, assume that $2 \in \tilde{S}_2$ and $e_4 \in E_2$ ($G_2 = (V_2, E_2)$ is as defined earlier). The edge e_1 belongs to $\tilde{E}_1$ or E_2, since e_1 and e_4 belong to the same K_4 or series-parallel component. Define $\tilde{G} = (\tilde{V}, \tilde{E}) = \tilde{G}_1 +_2 G_2$ and $\hat{G} = (\tilde{V}, \tilde{E} - \{e_1\})$, where e_4 is the edge identified in the 2-sum. Let

$$(3.14) \qquad \tilde{a}^i x \leq \tilde{a}_0^i, \qquad i = 1, 2$$

be the inequalities obtained as 2-sums of the inequalities in (3.12) indexed by $\{1, \ldots, r\} - \tilde{S}_2$. Each of inequalities (3.14) is facet-defining for $P(\tilde{G}_1)$. The inequalities

$$(3.15) \qquad a^{i2} x \leq a_0^{i2}, \qquad i = 1, 2$$

are the inequalities in (3.12) that are facet-defining for $P(G_2)$. Note that at least one of $a^{12}(e_4) \neq 0$ or $a^{22}(e_4) \neq 0$. Consider the inequalities

$$(3.16) \qquad \hat{a}^i x \leq \hat{a}_0^i, \qquad i = 1, 2$$

obtained as 2-sums of (3.14) and (3.15) on identifying edge e_4. By Proposition 3.3, each inequality in (3.16) is facet-defining for $P(\tilde{G})$. Define

$$(3.17) \qquad \hat{a}x \leq \hat{a}_0,$$

where $\hat{a}(e) = \hat{a}^1(e) + \hat{a}^2(e)$ for all $e \in \tilde{E} - \{e_1\}$ and $\hat{a}_0 = \hat{a}_0^1 + \hat{a}_0^2$. Note that $\hat{a}(e) = a(e)$ for all $e \in \bar{E} \cap \tilde{E}$. Furthermore, each partition satisfying (3.13) with equality is the extension of a partition satisfying (3.17) with equality. Thus, if inequality (3.17) is not facet-defining for $P(\hat{G})$, then (3.13) cannot be facet-defining for $P(\bar{G})$. If $\hat{a}(e) \neq 0$ for some edge $e \in E_2 - \{e_4\}$, then inequality (3.17) is not facet-defining for $P(\hat{G})$ by Proposition 3.7 (since $\hat{a}(\bar{e}) \neq 0$ and $\bar{e} \in \tilde{E}_1 - \{e_4\}$). Thus, inequality (3.13) is not facet-defining for $P(\bar{G})$ in this case.

Now consider the case where $\hat{a}(e) = 0$ for all edges $e \in E_2 - \{e_4\}$, i.e., $\hat{a}^1(e) + \hat{a}^2(e) = 0$ for all $e \in E_2 - \{e_4\}$. Define

$$E_2^i = \{e \in E_2 \mid a^{i2}(e) \neq 0\}, \qquad i = 1, 2.$$

Since G_2 is either a K_4 or a series-parallel graph, E_2^i is the subgraph of a cycle or a 3-wheel. The only cycle inequalities for which $\hat{a}_e^1 + \hat{a}_e^2 = 0$ for all $e \in E_2 - \{e_4\}$ is a triangle containing e_4, e_5, and e_6, where $a^{12}(e_4) = a^{22}(e_4) = -1$, $a^{12}(e_5) = 1$, $a^{22}(e_5) = -1$, $a^{12}(e_6) = -1$, $a^{22}(e_6) = 1$. If the two inequalities are 3-wheel inequalities, once again there must be a triangle as described above. Thus, there exist edges e_5 and e_6 such that $\{e_4, e_5, e_6\}$ define a triangle in G_2 with

$$\hat{a}_e^1 = -\hat{a}_e^2 = \begin{cases} 1 & \text{for } e = e_5, \\ -1 & \text{for } e = e_6, \\ 0 & \text{for } e \in E_2 - \{e_4, e_5, e_6\}. \end{cases}$$

Since each solution that satisfies (3.17) with equality has both u and v in the same subset, it satisfies $x(e_5) = x(e_6)$. Thus, inequality (3.17) is not facet-defining for $P(\tilde{G})$, since $P(\tilde{G})$ is full-dimensional. This implies that inequality (3.13) is not facet-defining for $P(G)$. The result thus follows. □

Proof of Theorem 3.1. Let $G = (V, E)$ be a 4-wheel free graph of minimum size that violates Theorem 3.1. G cannot be series-parallel or a K_4 by Theorem 2.1 and Proposition 3.1, respectively. Thus G must be the 2-sum (or $\bar{2}$-sum) of two 4-wheel free graphs $G_i = (V_i, E_i)$, $i = 1, 2$. First, consider the case where $G = G_1 \bar{+}_2 G_2$. Theorem 3.1 holds for G_1 and G_2 by the induction hypothesis. Thus, by Proposition 3.2, Theorem 3.1 also holds for G.

Now consider the case where $G = G_1 +_2 G_2$. Assume that the edges identified and deleted in the 2-sum are $e_i \in E_i$, $i = 1, 2$. Let $\bar{G} = (V, \bar{E})$ be the $\bar{2}$-sum of G_1 and G_2 obtained on identifying e_1 and e_2 into the edge $\bar{e}$. Thus, $\bar{E} = E \cup \{\bar{e}\}$. By assumption, the polytopes $P(G_1)$ and $P(G_2)$ are completely defined, as in Theorem 3.1. From Proposition 2.1, it follows that $P(\bar{G})$ is completely defined by inequalities of type (a), (b), and (c) in the statement of Theorem 3.1.

The polytope $P(G)$ is the projection of $P(\bar{G})$ onto the set R^E. Thus we must only show that the projection of inequalities of type (a), (b), or (c) (that completely define $P(\bar{G})$) gives inequalities in one of the three classes. We use the same techniques as in the proof of Proposition 2.3.

The matrix A is the column corresponding to the edge $\bar{e}$. The column has an entry of 0 or 1 for all inequalities of the form $x_e \leq 1$ and 0, 1 or -1 for all inequalities of the form (b) or (c). Let m be the total number of inequalities defining $P(\bar{G})$. Define the cone W, where

$$W = \{t \in R^m \mid tA \geq 0, t \geq 0\}.$$

As in the proof of Proposition 2.3, the extreme rays t of W are of the form (2.10) or (2.11).

Extreme rays of the type (2.10) are unit vectors and do not alter the inequality on projection. Extreme rays of type (2.11) add two inequalities, one of which has a coefficient of $+1$ in $\bar{e}$, and the other has a coefficient of -1 in $\bar{e}$. If one of the inequalities is $x(\bar{e}) \leq 1$ and the other an inequality from WI, then the composed inequality cannot be facet-defining by Proposition 3.6. If both inequalities are cycle inequalities whose support graphs share another edge besides $\bar{e}$, the composed inequality is not facet-defining as in the proof of Proposition 2.3. If one is an inequality from WI and the other is either a cycle inequality or an inequality from WI and the two support graphs share another edge besides $\bar{e}$, the composed inequality is not facet-defining by Proposition 3.10. If the two support graphs only share the edge $\bar{e}$, the composed inequality is once again from the class WI.

This shows that the projection of $P(\bar{G})$ onto the set $\{x \mid x(\bar{e}) = 0\}$ is completely defined by inequalities of type (a), (b), or (c). This contradicts our assumption that G is a 4-wheel free graph that violates Theorem 3.1. The result thus follows. □

Remark 3.1. Theorem 3.1 shows that, if G is a 4-wheel free graph, all inequalities defining $P(G)$ have coefficients of 0, 1, or -1 if the inequality has been scaled by the smallest integer to ensure that all coefficients and the right side are integer.

Acknowledgments. I thank the referees for a careful reading of the paper and several valuable suggestions that have resulted in a significantly improved version. In particular, the proof of Proposition 3.10 was incorrectly stated in the original version, as pointed out by an anonymous referee.

REFERENCES

[1] E. BALAS AND W. PULLEYBLANK, *The perfectly matchable subgraph polytope of a bipartite graph*, Networks, 13 (1983), pp. 495–516.

[2] M. W. BERN, E. L. LAWLER, AND A. L. WONG, *Linear time computation of optimal subgraphs of decomposable graphs*, J. Algorithms, 8 (1987), pp. 216–235.

[3] J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, Elsevier, New York, 1976.

[4] S. CHOPRA AND M. R. RAO, *The partition problem*, Math. Programming, 59 (1993), pp. 87–115.

[5] M. DEZA, M. GRÖTSCHEL, AND M. LAURENT, *Clique-Web Facets for Multicut Polytopes*, Report No. 186, Institut für Mathematik, Universität Augsburg, 1989.

[6] ———, *Complete Descriptions of Small Multicut Polytopes*, Report No. 217, Institut für Mathematik, Universität Augsburg, 1990.

[7] M. GRÖTSCHEL, L. LOVASZ, AND A. SCHRIJVER, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[8] M. GRÖTSCHEL AND Y. WAKABAYASHI, *Facets of the clique partitioning polytope*, Math. Programming, 47 (1990), pp. 367–387.

[9] P. D. SEYMOUR, *Decomposition of regular matroids*, J. Combin. Theory Ser. B, 28 (1980), pp. 305–359.

[10] J. A. WALD AND C. J. COLBOURN, *Steiner trees, partial 2-trees, and minimum IFI networks*, Networks, 13 (1983), pp. 159–167.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 32–40, February 1994

ON THE INTERPLAY BETWEEN INTERVAL DIMENSION AND DIMENSION*

S. FELSNER†, M. HABIB‡, AND R. H. MÖHRING†

Abstract. This paper investigates a transformation $P \twoheadrightarrow Q$ between partial orders P, Q that transforms the interval dimension of P to the dimension of Q, i.e., idim $(P) =$ dim (Q). Such a construction has been shown before in the context of Ferrer's dimension by Cogis [*Discrete Math*., 38 (1982), pp. 47–52]. The construction in this paper can be shown to be equivalent to his, but it has the advantage of (1) being purely order-theoretic, (2) providing a geometric interpretation of interval dimension similar to that of Ore [*Amer. Math. Soc. Colloq. Publ*., Vol. 38, 1962] for dimension, and (3) revealing several somewhat surprising connections to other order-theoretic results. For instance, the transformation $P \twoheadrightarrow Q$ can be seen as almost an inverse of the well-known split operation; it provides a theoretical background for the influence of edge subdivision on dimension (e.g., the results of Spinrad [*Order*, 5 (1989), pp. 143–147]) and interval dimension, and it turns out to be invariant with respect to changes of P that do not alter its comparability graph, thus also providing a simple new proof for the comparability invariance of interval dimension.

Key words. partial order, dimension, interval dimension, interval order, geometric representation of ordered sets

AMS subject classifications. 06A10, 68E10, 68C25

1. Introduction.

An *extension* Q of a partial order P is an order on the same elements that contains all the ordered pairs of P; i.e., $x < y$ in P implies $x < y$ in Q. A family $\{Q_1, \dots, Q_k\}$ of extensions of P is said to *realize* P or to be a *realizer* of P if and only if $P = Q_1 \cap \cdots \cap Q_k$; i.e., $x < y$ in P if and only if $x < y$ in Q_i for each i, $1 \leq i \leq k$. If we restrict the Q_i to belong to a special class of orders and seek for a minimum size realizer, we obtain a concept of dimension with respect to the special class.

A *linear extension* of P is an extension that is a chain. Dushnik and Miller [5] defined the *dimension* of a partial order P, denoted dim (P), as the smallest integer k for which there exist k linear extensions realizing P. Let the realizer $\{L_1, \dots, L_k\}$ of an order P with $|P| = n$ be made by the linear extensions $L_i = (x_{i_1} < x_{i_2} < \cdots < x_{i_n})$. With every $x \in P$, we then associate the vector $(x^1, \dots, x^k) \in \mathbf{R}^k$, where x^i gives the position (coordinate) of x in L_i. This mapping of the points of P to points of $\mathbf{R}^k$ embeds P into the componentwise ordering of $\mathbf{R}^k$. Ore [15] defined dim (P) as the minimum k such that P embeds into $\mathbf{R}^k$ in this way. The projections of such an embedding on each coordinate yield a realizer of P. The extensions of this realizer need not be linear extensions of P, but it is straightforward to transform them into linear extensions. Therefore, Ore dimension and Dushnik–Miller dimension are equivalent concepts.

A partial order P is an *interval order* if it can be represented by assigning to each element $x \in P$ an open interval $I_x = (a_x, b_x)$ of the real line, such that $x < y$ in P if and only if $b_x \leq a_y$. Such a collection of intervals is called an *interval representation* of P. Fishburn [7] characterized interval orders as those orders that do not contain a four-point subset forming two disjoint 2-chains, i.e., that contain no $2 + 2$. Interval dimension,

* Received by the editors September 3, 1991; accepted for publication (in revised form) January 6, 1992. This work was supported in part by the Deutsche Forschungsgemeinschaft and by the Deutscher Akademischer Austausch Dienst (DAAD) Project PROCOPE.

† Technische Universitat Berlin, Berlin, Germany.

‡ Laboratoire d'Informatique de Robotique de Microélectronique de Montpellier (LRIMM), Montpellier, France.

denoted idim(P), is defined by using interval extensions instead of linear extensions. Since linear orders are interval orders, we obtain the trivial inequality

$$(1) \qquad \operatorname{idim}(P) \leq \dim(P).$$

It is well known that interval orders of large dimension exist (see [16], [1], [9]). Hence, the gap between idim(P) and dim(P) may be arbitrarily large.

Let $\mathscr{I} = \{I_1, \ldots, I_k\}$ be an interval realizer of P and fix an open interval representation of each interval order I_j. Let (a_x^j, b_x^j) be the interval corresponding to $x \in P$ in the representation of I_j. We now define a *box embedding* of P to $\mathbf{R}^k$. With $x \in P$, we associate the box $\Pi_j\,(a_x^j, b_x^j) \subseteq \mathbf{R}^k$. Each of these boxes is uniquely determined by its *upper extreme corner* $u_x = (b_x^1, \ldots, b_x^k)$ and its *lower extreme corner* $l_x = (a_x^1, \ldots, a_x^k)$. Obviously, $x < y$ in P if and only if $u_x \leq l_y$ componentwise. The projections of a box embedding onto each coordinate yield an interval realizer, so the concepts of box embeddings and interval realizers are equivalent. For interval dimension, the box embeddings thus play the role of the above-mentioned point embeddings into $\mathbf{R}^k$ introduced by Ore for dimension.

A box embedding depends not only on the realizer $\mathscr{I}$ of P, but also on the representations of the I_j. Now we define the *partial order* $\mathscr{B}(\mathscr{I})$ *of extreme corners* associated with a box embedding or, equivalently, with an interval realizer $\mathscr{I}$ of P. The vertices of $\mathscr{B}(\mathscr{I})$ are the, at most $2n$ different, lower (respectively, upper) extreme corners of elements of P. The order relation of $\mathscr{B}(\mathscr{I})$ is given by the componentwise order in $\mathbf{R}^k$.

By definition, we have an embedding of $\mathscr{B}(\mathscr{I})$ in $\mathbf{R}^k$, so $\dim \mathscr{B}(\mathscr{I}) \leq k = \operatorname{idim}(P)$. The starting point of our investigations is the following question concerning the interplay between dimension and interval dimension:

$$\text{Is } \dim \mathscr{B}(\mathscr{I}) = \operatorname{idim}(P)?$$

In § 2 we define a transformation $P \to B(P)$ such that $\operatorname{idim}(P) = \dim B(P)$. We provide two interpretations of this transformation, a combinatorial one and a geometrical one. In the combinatorial interpretation, the elements of $B(P)$ are subsets of P. In the geometrical interpretation, $B(P)$ is the poset $\mathscr{B}(\mathscr{I}^*)$ of extreme corners of $\mathscr{I}^*$. Here $\mathscr{I}^*$ is a box embedding obtained from an arbitrary box embedding $\mathscr{I}$ by a normalizing procedure. From the proofs, we obtain an affirmative answer to the above question.

Section 3 investigates several consequences to and relations with other order-theoretic results. First, we study the transformation $P \to B(P)$ on special partial orders of height 1. In particular, we show that the standard example S_n of a n-dimensional order is an (almost) fixed point of the transformation $P \to B(P)$. Therefore, $\dim(S_n) = \operatorname{idim}(S_n)$.

Second, we investigate the relationship with the split operation. This has a surprising consequence for the iterated transformation $P \to B(P) \to B^2(P) \to \cdots B^k(P) \to \cdots$. For every n, there are partial orders P such that $0 \leq \dim(P) - \dim B^k(P) \leq 2$ for all $k \leq n$, but $\dim(P) - \dim B^{n+1}(P) \geq m$, where m is arbitrary.

Third, we relate the interval dimension of subdivisions of P to the dimension of P, thus providing a theoretical framework for the examples of Spinrad [17].

Finally, we show the comparability invariance of the transformation $P \to B(P)$, which, as a consequence, gives another proof that the interval dimension is a comparability invariant. Some remarks concerning the recognition-complexity of special classes of orders and graphs close the paper.

2. The main result. In the last section, we define the partial order $\mathscr{B}(\mathscr{I})$ of extreme corners associated with a box embedding of P in $\mathbf{R}^k$. With the next lemmas, we show that $\mathscr{B}(\mathscr{I})$ inherits some structure that is independent of the realizer $\mathscr{I}$ leading to the box embedding.

LEMMA 1. *Let $\mathscr{B}(\mathscr{I})$ be the partial order of extreme corners of a box representation of P.*

(a) *If the lower extreme corners of x and y are comparable in $\mathscr{B}(\mathscr{I})$, e.g., $l_x \leq l_y$, then the predecessor sets of x and y in P are ordered by inclusion, i.e.,* $\mathrm{Pred}_P(x) \subseteq \mathrm{Pred}_P(y)$.

(b) *If the upper extreme corners of x and y are comparable in $\mathscr{B}(\mathscr{I})$, e.g., $u_x \leq u_y$, then the successor sets of x and y in P are ordered by (reversed) inclusion, i.e.,* $\mathrm{Succ}_P(x) \supseteq \mathrm{Succ}_P(y)$.

(c) *If the lower extreme corner of x and the upper extreme corner of y are related by $l_x \leq u_y$, then* $\mathrm{Pred}_P(z) \supseteq \mathrm{Pred}_P(x)$ *for all* $z \in \mathrm{Succ}_P(y)$ *or, equivalently,* $\mathrm{Pred}_P(x) \subseteq \bigcap_{z \in \mathrm{Succ}_P(y)} \mathrm{Pred}_P(z)$.

(d) *If $u_x \leq l_y$, then* $\bigcap_{z \in \mathrm{Succ}_P(x)} \mathrm{Pred}_P(z) \subseteq \mathrm{Pred}_P(y)$.

Proof. (a) From $l_x \leq l_y$, we obtain $a^j_x \leq a^j_y$ for all j. Therefore, in each I_j, x has less predecessors than y; i.e., $\mathrm{Pred}_j(x) \subseteq \mathrm{Pred}_j(y)$. The claim now follows from $\mathrm{Pred}_P(x) = \bigcap_j \mathrm{Pred}_j(x)$ since the I_j realize P.

(b) The proof of this part is symmetric to part (a).

(c) From $l_x \leq u_y$, we have $a^j_x \leq b^j_y$. If $z \in \mathrm{Succ}(y)$, then necessarily $a^j_z \geq b^j_y \geq a^j_x$. Hence, $\mathrm{Pred}_j(z) \supseteq \mathrm{Pred}_j(x)$ for all j. The claim follows.

(d) From $u_x \leq l_y$, we immediately obtain $x \leq y$, i.e., $y \in \mathrm{Succ}(x)$; therefore, $\mathrm{Pred}(y) \supseteq \bigcap_{z \in \mathrm{Succ}(x)} \mathrm{Pred}(z)$. □

All statements except the conclusion part of (b) use only the sets $\mathrm{Pred}_P(x)$ and $\bigcap_{z \in \mathrm{Succ}_P(x)} \mathrm{Pred}_P(z)$. This irregularity is resolved with the next lemma.

LEMMA 2. $\bigcap_{z \in \mathrm{Succ}(x)} \mathrm{Pred}(z) \supseteq \bigcap_{z \in \mathrm{Succ}(y)} \mathrm{Pred}(z)$ *if and only if* $\mathrm{Succ}(x) \subseteq \mathrm{Succ}(y)$.

Proof. The "if" direction is trivial. We now prove the "only if" direction. Let $z \in \mathrm{Succ}(x)$ and note that $y \in \bigcap_{z \in \mathrm{Succ}(y)} \mathrm{Pred}(z)$. From the assumed inclusion, we obtain $y \in \mathrm{Pred}(z)$; hence $z \in \mathrm{Succ}(y)$. □

DEFINITION 1. With each vertex x of a partial order P, we associate the *lower set* $L(x) = \mathrm{Pred}_P(x)$ and the *upper set* $U(x) = \bigcap_{z \in \mathrm{Succ}_P(x)} \mathrm{Pred}_P(z)$. The case where $x \in \mathrm{Max}(P)$ is settled by the convention $U(x) = P$.

Define $B(P) = \{L(x), U(x) : x \in P\}$ ordered by setinclusion.

Note that this construction is, in fact, equivalent with Cogis's construction in the context of Ferrers dimension [2]. Cogis also uses $L(x)$ but replaces $U(x)$ by the equivalent set $\{z \in P : \mathrm{Succ}(x) \subseteq \mathrm{Succ}(z)\}$. He also proves Theorem 2, but in a different way and without the geometrical interpretation upon which our approach is based.

The preceding lemmas prove that $l_x \twoheadrightarrow L(x)$ and $u_x \twoheadrightarrow U(x)$ together form an order-preserving mapping from $B(\mathscr{I})$ to $B(P)$; hence

$$\mathrm{idim}(P) \geq \dim \mathscr{B}(\mathscr{I}) \geq \dim B(P). \tag{2}$$

To get more structure into interval realizers, we now introduce a procedure that transforms an interval extension $I = \{(a_x, b_x) : x \in P\}$ of P into its *P-normalization* $I^* = \{(a^*_x, b^*_x) : x \in P\}$.

In the first step of the P-normalization, we update left endpoints as follows:

$$a^*_x = \max\{b_z : z \in \mathrm{Pred}(x)\} \quad \text{if } x \text{ is not minimal,}$$

$$a^*_x = \min\{a_z : z \in \mathrm{Min}(P)\} \quad \text{if } x \text{ is minimal.}$$

In the second step, we update right endpoints as follows:

$$b^*_x = \min\{a^*_z : z \in \mathrm{Succ}(x)\} \quad \text{if } x \text{ is not maximal,}$$

$$b^*_x = \max\{b_z : z \in \mathrm{Max}(P)\} \quad \text{if } x \text{ is maximal.}$$

Note that the interval order I^* need not be isomorphic to I. In general, I^* is a suborder of I and a minimal interval extension of P if all the a_x, b_x were different.

If P is realized by $\mathcal{I} = \{I_1, \ldots, I_k\}$, then $\mathcal{I}^* = \{I_1^*, \ldots, I_k^*\}$ realize P as well. We call the box embedding corresponding to $\mathcal{I}^*$ the *normalized* box embedding of $\mathcal{I}$. For an example, see Fig. 1.

After normalizing, we have a realizer $\mathcal{I}^* = \{I_1^*, \ldots, I_k^*\}$ of P, interval representations (a_x^{j*}, b_x^{j*}) and an associated partial order of extreme corners $\mathcal{B}(\mathcal{I}^*) = \{l_x^*, u_x^* : x \in P\}$. The next theorem shows that the geometrically defined $\mathcal{B}(\mathcal{I}^*)$ and the combinatorially defined $B(P)$ are isomorphic.

THEOREM 1. *If $\mathcal{I}^*$ is a normalized realizer of P, then $\mathcal{B}(\mathcal{I}^*) = B(P)$.*

Proof. First, observe that both partial orders have a least element generated by $x \in \mathrm{Min}(P)$ as l_x^* and $L(x)$, respectively, and a greatest element generated by $x \in \mathrm{Max}(P)$ as u_x^* and $U(x)$, respectively.

Moreover, $l_x^* \twoheadrightarrow L(x)$ and $u_x^* \twoheadrightarrow U(x)$ defines an order-preserving mapping by the remarks preceeding (2). To show the converse, we distinguish the following four cases:

$U(x) \subseteq L(y)$. We know that $x \in U(x)$, so $x \in \mathrm{Pred}(y)$. Since I_j^* is an interval extension of P, we obtain $b_x^{j*} \le a_y^{j*}$ for all j. Hence $u_x^* \le l_y^*$;

$L(x) \subseteq L(y)$. Remember that $a_x^{j*} = \max\{b_z^j : z \in \mathrm{Pred}(x)\}$ and $a_y^{j*} = \max\{b_z^j : z \in \mathrm{Pred}(y)\}$. By assumption, $\mathrm{Pred}(x) \subseteq \mathrm{Pred}(y)$, so $a_x^{j*} \le a_y^{j*}$ and $l_x^* \le l_y^*$;

$U(x) \subseteq U(y)$. By Lemma 2, this is equivalent to $\mathrm{Succ}(x) \supseteq \mathrm{Succ}(y)$. Now $u_x^* \le u_y^*$ follows symmetrically with the second case;

$L(x) \subseteq U(y)$. Since I_j^* is normalized, there are $z_0 \in \mathrm{Pred}(x)$ and $z_1 \in \mathrm{Succ}(y)$ with $a_x^{j*} = b_{z_0}^j$ and $b_y^{j*} = a_{z_1}^{j*}$. The hypothesis provides $z_0 \le z_1$; hence $a_x^{j*} = b_{z_0}^j \le b_{z_0}^{j*} \le a_{z_1}^{j*} = b_y^{j*}$. The validity of this inequality for all j again gives $l_x^* \le u_y^*$. □

We are now ready to prove our main theorem about interval dimension and dimension.

THEOREM 2. *It holds that* $\dim B(P) = \mathrm{idim}(P)$.

Proof. As inequality (2), we already have obtained $\dim B(P) \le \mathrm{idim}(P)$. For the converse, we need two arguments. We first show that a linear extension L of $B(P)$ induces an interval extension I_L of P. Second, we prove that, if $L_1, \ldots, L_k$ is a realizer of $B(P)$, then the induced interval extensions $I_{L_1}, \ldots, I_{L_k}$ form an interval realizer of P.

Let $L = M_1, M_2, \ldots, M_r$ be a linear extension of $B(P)$. For each $x \in P$, there are $i, j \in \{1, \ldots, r\}$ such that $M_i = L(x)$ and $M_j = U(x)$. From $L(x) \subseteq U(x)$ and $x \notin L(x)$, $x \in U(x)$ we obtain that $i < j$. We can now associate with x a unique interval (a_x, b_x), which is defined to be (i, j). We now show that the interval order I_L induced

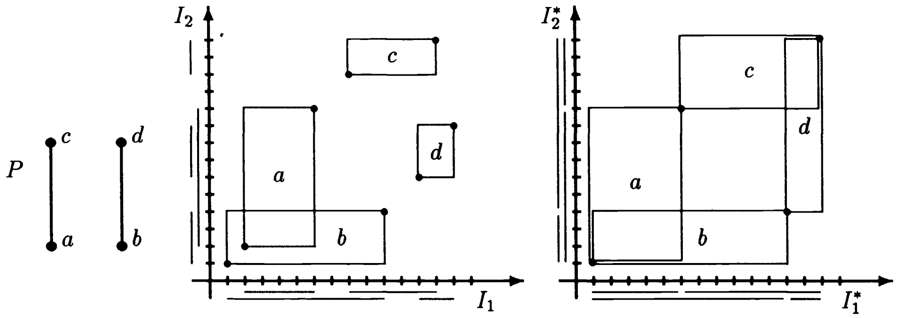


FIG. 1. *P, an interval realizer of P, and its normalization.*

by the interval representation $\{(a_x, b_x) : x \in P\}$ is an extension of P. If $x < y$ in P, then $U(x) \subseteq L(y)$, and thus, with $M_i = U(x)$ and $M_j = L(y)$, $b_x = i \leq j = a_y$, which implies $x < y$ in I_L.

Let $\{L_1, \ldots, L_k\}$ be a realizer for $B(P)$ and let $I_j = I_{L_j}$, for $1 \leq j \leq k$. The family $\{I_1, \ldots, I_k\}$ of interval extensions of P is an interval realizer if and only if all incomparabilities $x \| y$ of P are realized. If $x \| y$ in P, then $U(x) \not\subseteq L(y)$, since $x \in U(x)$ but $x \notin L(y)$. Therefore, $L(y)$ precedes $U(x)$ in some L_j, which gives $a_y^j < b_x^j$. The symmetric argument yields an i with $a_x^i < b_y^i$. Both inequalities together give $x \| y$ in $\bigcap_j I_j$. □

This theorem together with inequality (2) shows that $\dim \mathcal{B}(\mathcal{I})$ is independent of the interval realizer $\mathcal{I}$.

3. Consequences. In the previous section, we introduced the operation $P \twoheadrightarrow B(P)$ mapping partial orders to partial orders. We now investigate several connections to other order-theoretical topics and results. Note that $B(P)$ always has a greatest and a least element. We adopt the convention of calling orders with this property *bounded* and denote by $Q \twoheadrightarrow \widehat{Q}$ the bounding of a partial order Q; i.e., $\widehat{Q}$ is the order resulting from Q by adjoining a new greatest and a new least element.

We first look at the effect of the operator B applied to special classes of orders.

3.1. Transformation rules for special posets. Let S_n denote the standard poset of dimension n, i.e., the set of all one-element and $(n-1)$-subsets of an n-element set ordered by setinclusion. Then

$$B(S_n) = \widehat{S}_n. \tag{3}$$

Let C_r denote the r-cycle. The r-cycle is the three-dimensional poset on $2r$ elements $\{x_1, y_1, x_2, y_2, \ldots, x_r, y_r\}$ with comparabilities

$$x_1 < y_1,\ y_1 > x_2,\ x_2 < y_2, \ldots, x_r < y_r,\ y_r > x_1,$$

$$B(C_r) = \widehat{C}_r. \tag{4}$$

Let the Hasse diagram of T be a tree. The *truncation* of T, denoted by $\operatorname{tr}(T)$, is the induced tree on the nonleaf vertices of T. Then

$$B(T) = \widehat{\operatorname{tr}(T)}. \tag{5}$$

In particular, (3) shows that the standard example S_n of an n-dimensional order is (up to closures) a fixed point of the operation $P \twoheadrightarrow B(P)$, thus showing again that, for every $n \geq 3$, there are orders P with $\dim(P) = \operatorname{idim}(P) = n$.

3.2. *B* as an inverse of the split operation. We now turn to the natural question of whether, for every bounded partial order Q, there is some P with $Q = B(P)$. The next theorem answers this question affirmatively. Moreover, it turns out that the operation $P \twoheadrightarrow B(P)$ is an almost left inverse of the *split* operation S, which has applications in different branches of poset theory (see, e.g., [21], [8], [6]). The split $S[P]$ of a partial order P is the order of height 1 with minimal elements $\{x' : x \in P\}$ maximal elements $\{x'' : x \in P\}$ and ordered pairs $x' < y''$ if and only if $x \leq y$ in P.

THEOREM 3. *It holds that* $B(S[P]) = \widehat{P}$.

Proof. For $x \in P$, let $\operatorname{Pred}[x] = \operatorname{Pred}(x) \cup \{x\}$. Obviously, P is isomorphic to the setsystem $\{\operatorname{Pred}[x] : x \in P\}$ ordered by inclusion. We show that the elements of $B(S[P])$ are just the "primed" sets $\operatorname{Pred}[x]$, i.e., $(\operatorname{Pred}[x])' = \{x' : x \in \operatorname{Pred}[x]\}$, together with the greatest element $\{x', x'' : x \in P\}$ and the least element $\varnothing$. We have $L(x') = \varnothing$ and

$$U(x') = \bigcap_{z'' \in \operatorname{Succ}(x')} \operatorname{Pred}(z'') = \bigcap_{z \in \operatorname{Succ}[x]} (\operatorname{Pred}[z])'.$$

Since $x \in \mathrm{Succ}[x]$, $U(x') \subseteq (\mathrm{Pred}[x])'$. On the other hand, $\mathrm{Pred}[x] \subset \mathrm{Pred}[z]$ for $z \in \mathrm{Succ}(x)$. Together, this gives $U(x') = (\mathrm{Pred}[x])'$. Similarly, we obtain $L(x'') = \mathrm{Pred}(x'') = (\mathrm{Pred}[x])'$. Finally, $U(x'') = \{x', x'' : x \in P\}$ by definition since $\mathrm{Succ}(x'') = \varnothing$. □

It is easy to verify that

$$B(\widehat{P}) = \widehat{B(P)}. \tag{6}$$

So we may generalize the theorem to

$$B^n(S^n[P]) = \overset{\left.\begin{smallmatrix}\wedge\\ \vdots\\ \wedge\end{smallmatrix}\right\}n \text{ boundings}}{P} \tag{7}$$

Investigations on the effect of iterated splitting to the dimension [19] lead to the inequality

$$\dim P \leq \dim S^n[P] \leq 2 + \dim P \quad \text{for all } n. \tag{8}$$

As a consequence of (7) and (8), we obtain that, for every n, there are partial orders P such that

$$\dim P - \dim B^k(P) \leq 2 \quad \text{for all } k \leq n.$$

(Just take $P = S^n[Q]$ for some order Q.) If we choose Q, however, to be an m-dimensional interval order, we obtain a large difference in dimension with the next iteration, i.e.,

$$\dim P - \dim B^{n+1}(P) \geq m - 1. \tag{9}$$

3.3. The interval dimension of subdivisions. With the next theorem, we relate the interval dimension of subdivisions of P to the dimension of P. Spinrad [17] showed that the dimension of a subdivision of a partial order can be an arbitrary multiple of its dimension, thus answering Trotter's problem 4 in [22]. With our result, we establish a theoretical framework for his examples.

In this context, partial orders and their diagrams are regarded as directed graphs whose edges (x, y) correspond to ordered pairs and cover pairs $x \prec y$ of P, respectively. An edge (x, y) is *subdivided* by placing a new vertex z in the "middle" of the edge; i.e., (x, y) is replaced by (x, z) and (z, y). In the case of partial orders, we must then ensure transitivity; i.e., all edges (a, z) with $a \in \mathrm{Pred}[x]$ and (z, b) with $b \in \mathrm{Succ}[y]$ are also added.

The *complete diagram subdivision* $DS(P)$ is the subdivision of all edges of the diagram of P. The *complete subdivision* $CS(P)$ is the transitive closure of the subdivision of all the edges of P. Let E be any set of edges of P; we denote the order obtained by subdividing the edges of E with $\mathrm{Sub}(P, E)$. Since P is an induced suborder of each subdivision $\mathrm{Sub}(P, E)$ and since $\mathrm{Sub}(P, E)$ is an induced suborder of $CS(P)$, we obtain

$$\dim(P) \leq \dim \mathrm{Sub}(P, E) \leq \dim CS(P). \tag{10}$$

With the next theorem, we give an upper bound for idim $\mathrm{Sub}(P, E)$.

THEOREM 4. *We have that* idim $\mathrm{Sub}(P, E) \leq \mathrm{idim} DS(P) = \mathrm{idim} CS(P) = \dim(P)$.

Proof. Take any embedding of P into $\mathbf{R}^k$ with $k = \dim(P)$ and grow the points to obtain an embedding by "miniboxes." An interval embedding of a subdivision $\mathrm{Sub}(P, E)$ is then obtained by adding the box with lower extreme corner u_x and upper extreme corner l_y for the point z subdividing the edge $(x, y) \in E$. See Fig. 2. This gives idim $\mathrm{Sub}(P, E) \leq \dim(P)$, independent of the choice of E.

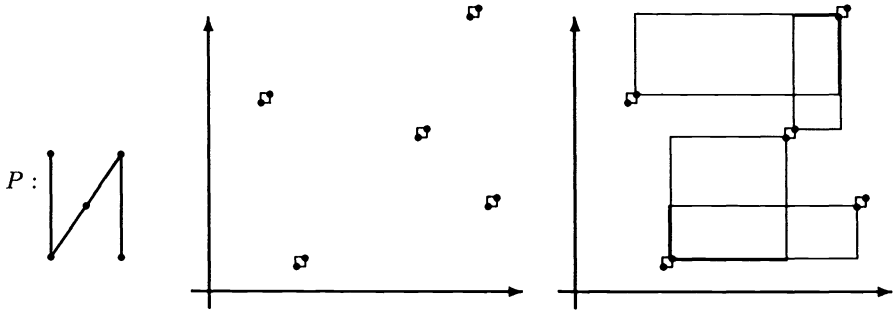


FIG. 2. *P, a minibox embedding of P, and the box embedding of $DS(P)$.*

To prove that $\operatorname{idim} DS(P) = \operatorname{idim} CS(P) = \dim(P)$, we show that P can be embedded in $B(DS(P))$. This gives $\dim(P) \le \dim B(DS(P))$. From Theorem 2, we know that $\operatorname{idim} DS(P) = \dim B(DS(P))$. Together, we obtain $\dim(P) \le \operatorname{idim} DS(P)$.

To show that P can be embedded in $B(DS(P))$, we apply the normalizing procedure to the box embedding $\mathscr{I}$ of $DS(P)$ constructed in the first paragraph of the proof. Because of the construction of the box embedding of $DS(P)$, the only changes that occur in normalization are shifts of the left endpoints of intervals corresponding to elements in $\operatorname{Min}(P)$ and the right endpoints of elements in $\operatorname{Max}(P)$. We then embed P into the lower extreme corners of the miniboxes of the normalized representation $\mathscr{I}^*$. This gives $\dim(P) = k = \operatorname{idim}(\mathscr{I}^*) = \operatorname{idim}(DS(P))$. □

Note that we obtained, in fact, a slightly stronger result: If a set E of edges of P contains the edges of the diagram, then $B(\operatorname{Sub}(P, E)) = VS(P)$, where $VS(P)$ denotes the *vertical split* of P, i.e., the order obtained from P by substituting each vertex by a 2-chain. In [20], a distinct proof for $\dim(P) = \operatorname{idim} VS(P)$ is given.

3.4. Comparability invariance. For the definition and basic facts on comparability invariance, see [10]. Let $\operatorname{Comp}(P)$ be the comparability graph of P. We show that $\operatorname{Comp}(B(P))$ is a comparability invariant of P in the sense that, if $\operatorname{Comp}(P) = \operatorname{Comp}(Q)$, then $\operatorname{Comp}(B(P)) = \operatorname{Comp}(B(Q))$. Together with Theorem 2 and the known fact that dimension is a comparability invariant, this gives an alternative proof of the comparability invariance of interval dimension in the finite case. The comparability invariance of interval dimension was first shown in [11].

A subset of elements A of P is called *autonomous* if the relation of elements in A to an element outside A is independent of the element of A. More formally, if, for any $x \notin A$, whenever $x < a$, $x > a$, or $x \| a$ holds for some $a \in A$, then it holds for all $a \in A$. If A is an autonomous subset of P, then $\operatorname{Pred}(A)$ denotes the predecessors outside A of any and hence every element of A.

THEOREM 5. *$\operatorname{Comp}(B(P))$ is a comparability invariant of P.*

Proof. Let A be an autonomous subset of P. It is enough to show (see, e.g., [11]) that $\operatorname{Comp}(B(P)) = \operatorname{Comp}(B(P^{A}_{A^d}))$, where $P^{A}_{A^d}$ denotes the order resulting from substituting A by its dual A^d into P.

Note first that $B(A) = \{L(a), U(a) : a \in A\}$ is a closed suborder of $B(P)$. Let $\widetilde{B(A)}$ be $B(A)$ without its greatest element $1_{B(A)}$ and its least element $0_{B(A)}$. Our claim is that $\widetilde{B(A)}$ is autonomous in $B(P)$. To see this, observe first that, for each $a \in A$, we can decompose $\operatorname{Pred}(a)$ into $\operatorname{Pred}(a) = \operatorname{Pred}(A) \cup \operatorname{Pred}_A(a)$; hence, the same is valid for all elements of $\widetilde{B(A)}$. On the other hand, the elements of $B(P) \backslash B(A)$ either contain all of A, or their intersection with A is empty. Now, if $M \in B(P) \backslash B(A)$ contains all of

A, then it also contains Pred(A), and M is above all sets in $B(\tilde{A})$. If $M \subseteq$ Pred(A), then M is below all sets in $B(\tilde{A})$. In all the other cases, M is unrelated to all of $B(\tilde{A})$. This gives the claim.

To settle the theorem, we need the fact that $B(A^d)$ and $B(A)^d$ are isomorphic orders. To see this, consider a normalized box embedding of A in $\mathbf{R}^k$. Its extreme corners are an embedding of $B(A)$ into $\mathbf{R}^k$. Flip the embedding, i.e., reverse the relations; this gives an embedding of A^d, and the extreme corners form an embedding of $B(A)^d$. □

As we have seen, autonomous sets in P induce autonomous sets in $B(P)$. The converse, however, is far from being true. Take as P a prime interval order; then $B(P)$ is a chain. Hence P has none, but $B(P)$ has $\binom{|B(P)|}{2} - 1$ nontrivial autonomous sets.

3.5. Remarks on related results and computational complexity. The transformation $P \twoheadrightarrow B(P)$ obviously can be carried out in polynomial time. The degree of the polynomial of this reduction is not important if we want to decide if P has interval dimension ≥ 3, since the poset dimension and interval dimension problems have been shown to be NP-complete for $k \geq 3$ by Yannakakis [23]. It is, however, a crucial point if we want to decide if idim$(P) \leq 2$.

Dagan, Golumbic, and Pinter [4] addressed questions about the comparability invariance of interval dimension and (fast) recognition of interval dimension at most 2. As remarked before, the first problem has been settled affirmatively first in [11]. The recognition problem has been solved independently by several authors. Habib and Möhring [12] used the subposet of $B(P)$ defined by $B'(P) = \{L(x) : x \in P\}$, together with an algorithm for transitive orientation with side constraints, to derive an $O(n \cdot n^a)$ algorithm where $O(n^a)$ is the best-known time for matrix multiplication. Currently a is approximately 2.37.

Cheah [3] proposed an algorithm in complexity $O(n^3)$. The same complexity is claimed by Langley [13]. Langley calls the poset $B(P)$ of the present paper the "predecessor-successor order" of P and shows without the aid of the geometric construction that finding an interval realizer of P is equivalent to finding a linear realizer of $B(P)$.

Spinrad [18] showed that recognition of two-dimensional orders only requires $O(n^2)$ operations. Therefore, the bottleneck with the $P \twoheadrightarrow B(P)$ approach for the recognition of interval dimension 2 is the computation of the order $B(P)$. With a careful implementation, $B(P)$ can be computed in $O(n^a)$, where again $O(n^a)$ is the best-known time for matrix multiplication.

Ma and Spinrad [14] found an approach that allows us to avoid matrix multiplication and leads to a complexity bound of $O(n^2)$. Since this is the best-known result, we outline the ideas behind their algorithm.

First, they construct the open split $S(P)$ of the partial order P, for which they want to decide if idim$(P) \leq 2$. The elements of the open split $S(P)$ are the same as of the split $S[P]$ defined above; the ordered pairs of $S(P)$ are, however, given by the irreflexive relation defining P, i.e., $x' < y''$ in $S(P)$ if and only if $x < y$ in P. They prove that the co-chain covering number of $S(P)$ equals the interval dimension of P. A theorem of Yannakakis [23] shows that the cochain covering number of $S(P)$ and the interval dimension of $S(P)$ coincide. Hence, idim$(P) =$ idim$(S(P))$, and we can reduce attention to the recognition of interval dimension 2 for bipartite orders, i.e., to orders of height 1.

The transformation of the interval dimension 2 problem for bipartite orders to the dimension 2 problem is done in two steps. In the first step, it is checked as to whether the bipartite order Q has a chordal bipartite comparability graph. A chordal bipartite graph is a bipartite graph without any induced cycle C_n, $n \geq 6$. If Comp(Q) is not chordal bipartite, then Q must contain a crown, and therefore idim$(Q) \geq 3$. Otherwise, the second step is started. In this step, Q is transformed into an order P_Q, which can be obtained by contracting autonomous chains in Stack(Q) and hence has the same dimension as Stack(Q). The stack operation was introduced by Trotter [21]. He proved that for bipartite posets dim Stack(Q) = idim(Q). As the construction of $B(Q)$, the construction of P_Q and Stack(Q) requires information about the containment relation of neighborhoods in Q. For chordal bipartite graphs, this information can be computed in $O(n^2)$ using a technique called doubly lexical ordering. Since after passing the first step we know that Q is chordal bipartite, we can construct P_Q and apply Spinrad's dimension 2 algorithm, all in $O(n^2)$.

Acknowledgment. We are grateful to an anonymous referee for his fruitful suggestions, which improved the presentation of the paper.

REFERENCES

[1] K. B. Bogart, I. Rabinovich, and W. T. Trotter, *A bound on the dimension of interval orders*, J. Combin. Theory Ser. A, 21 (1976), pp. 319–328.

[2] O. Cogis, *On the Ferrers dimension of a digraph*, Discrete Math., 38 (1982), pp. 47–52.

[3] F. Cheah, *A Recognition Algorithm for* II-*Graph*, Department of Computer Science, Univ. of Toronto, Toronto, Ontario, Canada, Tech. Report 246, 1990.

[4] ,I. Dagan, M. C. Golumbic, and R. Y. Pinter, *Trapezoid graphs and their coloring*, Discrete Appl. Math., 21 (1988), pp. 35–46.

[5] B. Dushnik and E. Miller, *Partially ordered sets*, Amer. J. Math., 63 (1941), pp. 600–610.

[6] S. Felsner, *Orthogonal structures in directed graphs*, J. Combin. Theory Ser. B, to appear.

[7] P. C. Fishburn, *Interval Orders and Interval Graphs*, John Wiley, New York, 1985.

[8] A. Frank, *On chain and antichain families of a partially ordered set*, J. Combin. Theory Ser. B, 29 (1980), pp. 176–184.

[9] Z. Füredi, V. Rödl,P. Hajnal, and W. T. Trotter, *Interval Orders and Shift Graphs*, preprint, 1991.

[10] M. Habib, *Comparability invariants*, Ann. Discrete Math., 23 (1984), pp. 331–386.

[11] M. Habib, D. Kelly, and R. H. Möhring, *Interval dimension is a comparability invariant*, Discrete Math., 88 (1991), pp. 211–229.

[12] M. Habib and R. H. Möhring, *Recognition of Partial Orders with Interval Dimension Two via Transitive Orientation with Side Constraints*, preprint 244, Technische Universitat, Berlin, 1990.

[13] L. J. Langley, *A Recognition Algorithm for Orders of Interval Dimension Two*, preprint, Dartmouth College, Hanover, NH, 1991.

[14] T. Ma and J. Spinrad, *An $O(n^2)$ time algorithm for the* 2-*chain cover problem and related problems*, in Proc. 2nd Ann. ACM-SIAM Sympos. on Discr. Alg., San Francisco, CA, 1991, pp. 363–372.

[15] O. Ore, *Theory of graphs*, Amer. Math. Soc. Colloq. Publ., Vol. 38, 1962.

[16] I. Rabinovich, *A note on the dimension of interval orders*, J. Combin. Theory Ser. A, 25 (1972), pp. 68–71.

[17] J. Spinrad, *Edge subdivision and dimension*, Order, 5 (1989), pp. 143–147.

[18] ———, *On comparability and permutation graphs*, SIAM J. Comput., 14 (1985), pp. 658–670.

[19] W. T. Trotter and J. I. Moore, *The dimension of planar posets*, J. Combin. Theory Ser. B, 22 (1977), pp. 54–67.

[20] ———, *Characterization problems for graphs, partially ordered sets, lattices and families of sets*, Discrete Math., 16 (1976), pp. 361–381.

[21] W. T. Trotter, *Stacks and splits of partially ordered sets*, Discrete Math., 35 (1981), pp. 229–256.

[22] ———, *Problems and conjectures in combinatorial theory of ordered sets*, Ann. Discrete Math., 41 (1989), pp. 401–416.

[23] M. Yannakakis, *The complexity of the partial order dimension problem*, SIAM J. Alg. Discrete Meth., 3 (1982), pp. 351–381.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 41-47, February 1994

THE PRIVATE NEIGHBOR CUBE*

MICHAEL FELLOWS†, GERD FRICKE‡, STEPHEN HEDETNIEMI§, AND DAVID JACOBS§

Abstract. Let S be a set of vertices in a graph $G = (V, E)$. The authors state that a vertex u in S has a private neighbor (relative to S) if either u is not adjacent to any vertex in S or u is adjacent to a vertex w that is not adjacent to any other vertex in S. Based on the notion of private neighbors, a set of eight graph theoretic parameters can be defined whose inequality relationships can be described by a three-dimensional cube. Most of these parameters have already been studied independently. This paper unifies this study and helps to form a cohesive theory of private neighbors in graphs. Theoretical and algorithmic properties of this private neighbor cube are investigated, and many open questions are raised.

Key words. NP-complete, irredundance, domination, independence, graph

AMS subject classifications. 05C, 68Q

1. Introduction.

Let $G = (V, E)$ be a graph and let v be a vertex in V. The *open neighborhood of* v is the set of vertices $N(v) = \{u \mid u \text{ is adjacent to } v\}$. The *closed neighborhood of* v is the set of vertices $N[v] = N(v) \cup \{v\}$. Let S be a subset of the vertices in V. The *open neighborhood of* S is the union $N(S) = \cup N(v)$ over all $v \in S$, and the *closed neighborhood of* S, $N[S] = \cup N[v]$ over all $v \in S$.

Let S be a set of vertices. We call a function p: $S \rightarrow V$ a *private neighbor function* on S if, for all $s \in S$, $p(s) \in N[s] - N(S - \{s\})$. We say that $p(s)$ is the *private neighbor* of s. It follows from this definition that p must be $1 - 1$. By placing various restrictions on p, we obtain several properties for vertex sets. Note that, for any $s \in S$, exactly one of the following three conditions holds for any private neighbor function p: (i) $p(s) = s$, (ii) $p(s) \in V - S$, or (iii) $p(s) \neq s$ and $p(s) \in S$. Let us define the following three predicates by forming the negation of these conditions:

$$Q(s) \equiv p(s) \neq s,$$

$$R(s) \equiv p(s) \in S,$$

$$P(s) \equiv p(s) = s \quad \text{or} \quad p(s) \in V - S.$$

Given a private neighbor function p on S, let us say that p has the following:

type 000	if for every $s \in S$,	$Q(s)$ and $R(s)$ and $P(s)$,
" 001	"	$Q(s)$ and $R(s)$,
" 010	"	$Q(s)$ and $P(s)$,
" 011	"	$Q(s)$,
" 100	"	$R(s)$ and $P(s)$,
" 101	"	$R(s)$,
" 110	"	$P(s)$, and
" 111	"	no restriction holds.

Finally, we say that a set S is a *type t set* if there exists a private neighbor function on S of type t. Note that, if S is a type $t = b_1b_2b_3$ set, then S is also a type t' set, where $t' = b_1b_2b_3 \vee 001$, $b_1b_2b_3 \vee 010$, or $b_1b_2b_3 \vee 100$. For example, any type 010 set also

* Received by the editors May 8, 1991; accepted for publication (in revised form) November 30, 1992.

† Department of Computer Science, P.O. Box 3055, University of Victoria, British Columbia, Canada V8W 3P-6.

‡ Department of Mathematics, Wright State University, Dayton, Ohio 45435-0001.

§ Department of Computer Science, Clemson University, Clemson, South Carolina 29634-1906.

has types 110, 011, and 111. We can restate this new terminology in more familiar graph theoretic terms, as follows.

Type 000. Each vertex in S must have a private neighbor; however, it cannot be itself, it cannot be outside of S, and it cannot be inside of S. Clearly, the *empty set*, $\varnothing$, is the only such set a graph G can have.

Type 001. Each vertex in S must have a private neighbor inside S other than itself. We see that the subgraph $\langle S \rangle$ induced by such a set consists of disjoint copies of the complete graph K_2 on two vertices. The set of edges in such an induced subgraph has been called an *induced matching* by Cameron [1] and a *strong matching* by Golumbic and Laskar [8]. (The latter appear to have discovered this notion independently.) We call such a set of vertices a *strong matching* set.

Type 010. Each vertex in S must have a private neighbor outside of the set S. We call such a set an *open irredundant set*; it satisfies the condition that, for every vertex $u \in S$, $N(u) - N[S - u] \neq \varnothing$. These sets have been studied by Farley and Schacham [5].

Type 011. Each vertex in S must have a private neighbor other than itself, either inside S or outside S. We call such a set an *open-open irredundant set*; it satisfies the condition that, for every vertex $u \in S$, $N(u) - N(S - u) \neq \varnothing$. These sets have been studied by Farley and Schacham [5] and Farley and Proskurowski [4].

Type 100. Each vertex in S must have itself as a private neighbor. It is easy to see that such a set is an *independent set* (that is, no two vertices in S are adjacent). Conversely, an independent set is a type 100 set.

Type 101. Each vertex in S must have either itself or a vertex inside S as a private neighbor. Such a set has been called a 1-*dependent set* by Fink and Jacobson [6]. In such a set S, the maximum degree of any vertex in $\langle S \rangle$ is less than or equal to 1, i.e., the subgraph induced by S consists of disjoint copies of K_1's and K_2's.

Type 110. Each vertex in S must have either itself or a vertex outside of S as a private neighbor. Such a set is called an *irredundant set*; it satisfies the condition that, for every vertex $u \in S$, $N[u] - N[S - u] \neq \varnothing$. Irredundant sets were first defined by Cockayne, Hedetniemi, and Miller in 1978 [3]. For a survey of results on irredundant sets in graphs, see Hedetniemi, Laskar, and Pfaff [12].

Type 111. A set S has a private neighbor function, but no further restriction is imposed. We call these *closed-open irredundant sets*; they satisfy the condition that, for every vertex $u \in S$, $N[u] - N(S - u) \neq \varnothing$. These sets were mentioned by Farley and Schacham [5].

In this paper, we are interested in studying, for each t, the *maximum cardinality* of a set of vertices having type t. These parameters are shown in Fig. 1.

Type	Maximum cardinality	Parameter
000	**0**	0
001	$\beta^*(G)$	Strong matching number
010	OIR(G)	Open irredundance number
011	OOIR(G)	Open-open irredundance number
100	$\beta_0(G)$	(Vertex) independence number
101	$\beta^1(G)$	1-dependence number
110	IR(G)	Irredundance number
111	COIR(G)	Closed-open irredundance number

FIG. 1. *The parameters of the private neighbor cube.*

The definitions of the eight parameters in Fig. 1 immediately give rise to 12 inequalities, as illustrated by the eight corners of the cube shown in Fig. 2, where the magnitudes of the parameters decrease in moving from right-to-left, back-to-front, and top-to-bottom. To place these parameters in a better perspective, we add the following well-known inequality chain:

$$\mathrm{ir} \leq \gamma \leq \mathrm{i} \leq \beta_0 \leq \Gamma \leq \mathrm{IR}.$$

(This was first observed by Cockayne, Hedetniemi, and Miller [3].) For completeness, we define these parameters. A set S of vertices in a graph $G = (V, E)$ is *dominating* if $N[S] = V$. The minimum and maximum cardinalities of a minimal dominating set in a graph G are denoted $\gamma(G)$ and $\Gamma(G)$, respectively; the minimum and maximum cardinalities of a maximal independent set in a graph G are denoted $i(G)$ and $\beta_0(G)$, respectively; and the minimum and maximum cardinalities of a maximal irredundant set in a graph G are denoted $\mathrm{ir}(G)$ and $\mathrm{IR}(G)$, respectively.

It is well known that, for arbitrary graphs G, $\beta_0(G) \leq \Gamma(G) \leq \mathrm{IR}(G)$. However, for bipartite graphs, Cockayne et al. have shown that $\beta_0(G) = \Gamma(G) = \mathrm{IR}(G)$ [2]. Similar results have been obtained for a variety of other interesting classes of graphs (cf. [3], [13], [14]). Golumbic and Laskar [8] showed that, for any bipartite graph G, $\beta^*(G) = \mathrm{OOIR}(G)$ and $\beta^1(G) = \mathrm{COIR}(G)$. Using results in [11], we can show that, for any tree $\mathrm{T} = (V, E)$, $\mathrm{IR}(T) + \mathrm{OIR}(T) = |V|$.

The goal of this paper is to establish the NP-completeness for the decision problems corresponding to the seven nontrivial parameters in Fig. 1. The problems for strong matching number and vertex independence, however, are already known to be NP-complete [1], [7]. Theorems 1–5 cover the remaining five parameters.

2. NP-completeness results. The oldest of these parameters, whose complexity is unknown, is perhaps the irredundance number, whose NP-completeness has remained open since its definition in 1978. Theorem 1 answers this question.

Let $G = (V, E)$ be an arbitrary graph. The *trestled graph of index* k, $T_k(G)$, is the graph obtained from G by adding k copies of K_2 to each edge uv of G and joining u and v to the endvertices of each K_2, respectively (cf. Fig. 3).

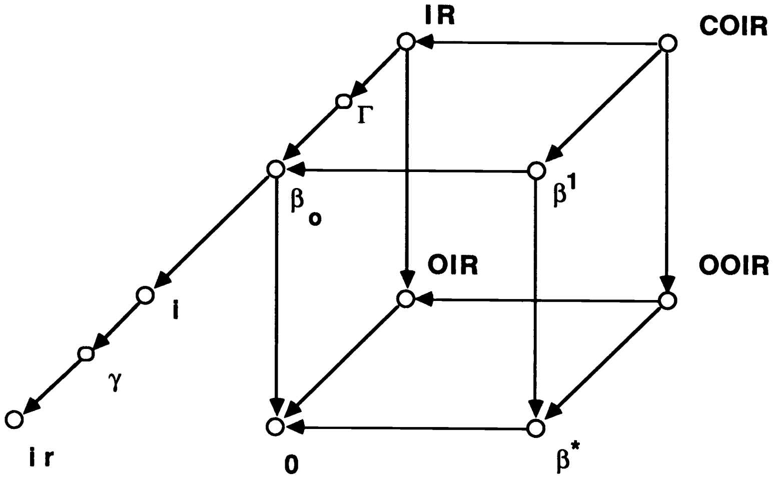


FIG. 2. *The private neighbor cube.*

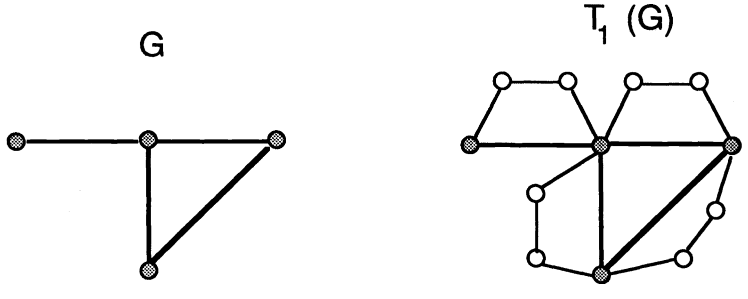


FIG. 3. *A trestled graph of index $k = 1$.*

LEMMA 1. *Let $G = (V, E)$ be any graph and let $T_2(G) = (V', E')$ be the corresponding trestled graph of index 2. For any irredundant set S' in $T_2(G)$, there exists an independent set S^* in $T_2(G)$ with $|S'| = |S^*|$.*

Proof. Let S' be an irredundant set in $T_2(G)$. Initially, let $S^* = S'$. Assume first that $S^* \cap V$ is not an independent set of vertices (cf. Fig. 4); i.e., there exist two adjacent vertices u and v in $S^* \cap V$. Let $P(u, v) = \{u', u'', v', v''\}$ in Fig. 4.

If u, v are in S^*, then $P(u, v) \cap S^* = \emptyset$ because S^* is irredundant. Therefore, $S^* = S^* - \{u, v\} \cup \{u', u''\}$ is an irredundant set having at least one less adjacent pair of vertices in V. Clearly, we can repeat this process for every pair of adjacent vertices u and v in $S^* \cap V$ until $S^* \cap V$ is an independent set.

Now consider $S^* \cap (V' - V)$. If this is not an independent set of vertices, then we can assume without loss of generality that there exists an adjacent pair u', v' in S^* (cf. Fig. 5). In this case, u, v, u'', and v'' are not in S^* because S^* is an irredundant set. Therefore, the set $S^* = S^* - \{v'\} \cup \{u''\}$ is irredundant and has exactly one less adjacent pair of vertices. Again, we can repeat this substitution process until $S^* \cap (V' - V)$ is an independent set.

Finally, consider the set S^*. If S^* is not an independent set, then, by the previous substitutions, there must exist a vertex u in V and a corresponding adjacent vertex u' in $V' - V$ that are both in S^* (cf. Fig. 6).

Consider the set of all vertices of the form u' or u'' in $S^* \cap (V' - V)$ that are adjacent to u. For each such u' or u'', the corresponding v' or v'' is not in S^* because by the previous substitutions, $S^* \cap (V' - V)$ is independent. Therefore, we can interchange all such u' and u'' vertices in $S^* \cap (V' - V)$ with corresponding vertices v' and v'' (leaving the vertex u in S^*). We can see that the resulting set S^* is irredundant and has fewer

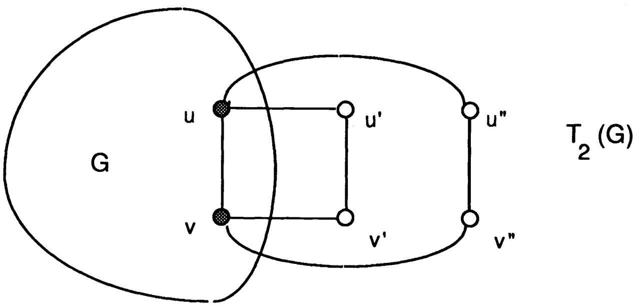


FIG. 4. *$S^* \cap V$ is not an independent set.*

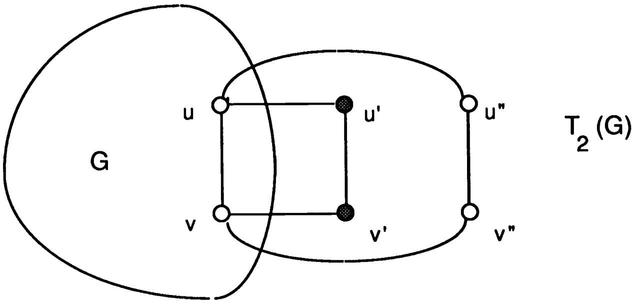


FIG. 5. $S^* \cap (V' - V)$ *is not an independent set.*

adjacent pairs than before. Note that the set $V(u)$ for all such v' and v'' vertices that we added to S^* is an independent set. Note further that no corresponding vertex v is in S^* (by the previous substitutions, $S^* \cap V$ is now an independent set). Therefore, vertex u is now independent in S^*. By repeating this substitution process of interchanging u' and u'' vertices for corresponding v' and v'' vertices, we can ultimately produce a set S^* that is independent and satisfies $|S^*| = |S'|$.

LEMMA 2. *A graph $G = (V, E)$ with $|E| = m$ edges has an independent set of size $\geq k$ if and only if the trestled graph of index 2, $T_2(G)$, has an irredundant set of size $\geq k + 2m$.*

Proof. Let S be an independent set of size $\geq k$ in G. Consider an arbitrary edge uv in E. If u and v are not in S, then $S \cup \{u', u''\}$ is an independent set. If u is in S and v is not in S then $S \cup \{v', v''\}$ is an independent set. Thus, for each edge uv in E, we can add two additional vertices to S to produce another independent set. As a result, $T_2(G)$ has an independent set of size $\geq k + 2m$ and hence an irredundant set of size $\geq k + 2m$ (since every independent set is irredundant).

Conversely, let S be an irredundant set of size $\geq k + 2m$ in $T_2(G)$. By Lemma 1, $T_2(G)$ has an independent set, say S^*, of size $\geq k + 2m$. Obviously, no independent set in $T_2(G)$ can contain more than two vertices in $P(u, v)$ for any edge uv

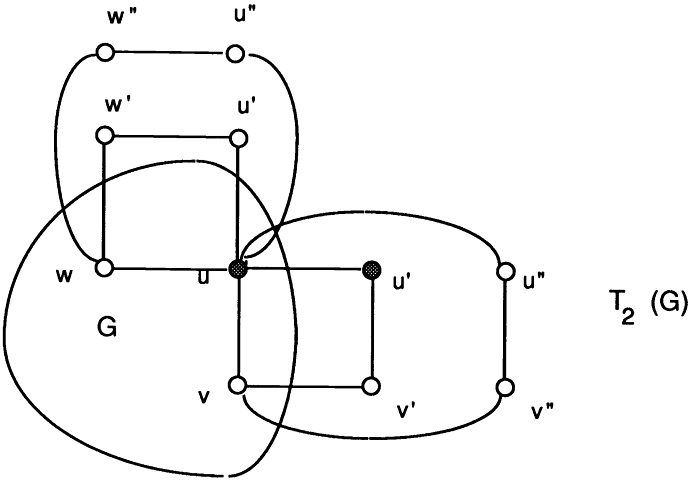


FIG. 6. S^* *is not an independent set.*

in E. Thus $|S^* \cap (V' - V)| \leq 2m$, and $|S^* \cap V| \geq k$, since $k + 2m \leq |S^*| = |S^* \cap (V' - V)| + |S^* \cap V|$.

We now formulate the following decision problem concerning IR(G):

IRREDUNDANT SET

INSTANCE: A graph $G = (V, E)$ and a positive integer k.

QUESTION: Does G have an irredundant set of size $\geq k$?

THEOREM 1. *The decision problem* IRREDUNDANT SET *is* NP-*complete.*

Proof. Clearly IRREDUNDANT SET is in NP. Its NP-completeness follows from Lemma 2 and the fact that INDEPENDENT SET is NP-complete [7].

Without proof, we mention that the decision problems corresponding to $\beta_0(G)$, $\Gamma(G)$, and IR(G) are all NP-complete for graphs having maximum vertex degree 5, trestled graphs of index k for any $k \geq 1$, planar graphs, and triangle-free graphs.

Given a graph $G = (V, E)$, we define the graph $G^+ = (V \cup V', E \cup E')$, called the *corona* $G \circ K_1$ (cf. Harary [11, p. 167]), to be the graph obtained from G by adding an adjacent endvertex to each vertex in G; i.e., for each vertex u in V, we add a new vertex u' and the edge uu'. It is easy to show the following lemma.

LEMMA 3. *For any graph* $G = (V, E)$ *with* $|V| = n$, $\beta_0(G) + n = \beta^1(G^+)$.

From Lemma 3 and the NP-completeness of INDEPENDENT SET, we obtain the following theorem.

THEOREM 2. *The decision problem for* β^1 *is* NP-*complete.*

We can also show that the decision problem for β^1 is NP-complete for both planar graphs and triangle-free graphs. In [1] Cameron showed that the decision problem corresponding to the parameter $\beta^*(G)$ is NP-complete, even for bipartite graphs. Without proof, we mention that this problem is NP-complete for triangle-free graphs and planar graphs, as well.

To solve the NP-completeness of the decision problem corresponding to the parameter COIR(G), all we need is the following lemma, which again uses the corona construction, and whose proof we omit.

LEMMA 4. *For any graph* G, $\beta^1(G^+) = \text{COIR}(G^+)$.

From Lemma 4 and Theorem 2, we now have the following theorem.

THEOREM 3. *The decision problem for* COIR *is* NP-*complete.*

Like previous parameters, COIR remains NP-complete for planar graphs and for triangle-free graphs. The NP-completeness questions are now answered for the decision problems associated with five of the parameters at the corners of the private neighbor cube: $\beta_0(G)$, $\beta^1(G)$, $\beta^*(G)$, IR(G), and COIR(G). The remaining NP-completeness questions for the parameters OIR(G) and OOIR(G) can be answered using matrix techniques.

THEOREM 4. *The decision problem for* OOIR *is* NP-*complete.*

Proof (sketch). Let G be a graph having closed neighborhood matrix $N(G)$. Now construct the graph H whose adjacency matrix is the following $2n \times 2n$ matrix:

$$\left[\begin{array}{c|c} 0 & N(G) \\ \hline N(G) & 0 \end{array}\right].$$

We can verify that $2(\text{IR}(G)) = \text{OOIR}(H)$.

Our construction shows that the decision problem for OOIR is NP-complete for bipartite graphs. It is also NP-complete for graphs having vertex degree at most 6.

THEOREM 5. *The decision problem for* OIR *is* NP-*complete.*

Proof (sketch). Given a graph G having n vertices, form a matrix M of size $2n$ similar to the one in the previous proof. However, this time, place 1's in the northwest and southeast quadrants. Let H be the graph having closed neighborhood matrix M. We can show that, if $\mathrm{IR}(G) > 2$, then $\mathrm{IR}(G) = \mathrm{OIR}(H)$.

3. Open problems. There are many open problems concerning the existence of polynomial time algorithms for computing the values of the parameters on the private neighbor cube for restricted classes of input graphs. The computation of the independence number $\beta_0(G)$ for various classes of graphs has been fairly well studied (cf. [15]). However, this is the only parameter for which this is the case. The value of $\beta_0(G)$ can be computed in linear time for trees (cf. [8]), in polynomial time for bipartite graphs (cf. [7]), and in polynomial time for families of graphs with bounded treewidth (cf. [15]), to name just a few results of this type for vertex independence.

The value of $\mathrm{IR}(G)$ can be computed in linear time for trees and polynomial time for bipartite graphs, because $\beta_0(G) = \mathrm{IR}(G)$ for bipartite graphs. However, we know of no other class of graphs for which $\mathrm{IR}(G)$ may be computed efficiently.

A linear time algorithm for computing $\mathrm{OIR}(T)$ for any tree T has been constructed by Farley and Proskurowski [4]. However, due to a result of Hedetniemi, $\mathrm{OIR}(T) = \beta^1(T)$ for any tree T, and $\mathrm{OIR}(T)$ can also be computed by any linear time matching algorithm for trees (cf. [9]). We know of no other $\mathrm{OIR}(G)$ algorithm for any other class of graphs.

Golumbic and Laskar [8] have shown that $\beta^*(G) = \mathrm{OOIR}(G)$ for bipartite graphs and circular arc graphs. In that paper, they also showed that this (combined) value can be computed in polynomial time for circular arc graphs. We know of no other algorithms for computing the values of $\beta^*(G)$ or $\mathrm{OOIR}(G)$ for any other classes of graphs.

Finally, Golumbic and Laskar [8] have shown that $\beta^1(G) = \mathrm{COIR}(G)$ for bipartite graphs. The only algorithmic result that we know for either of these two parameters is a linear time algorithm for computing this value for arbitrary trees by Hedetniemi.

REFERENCES

[1] K. Cameron, *Induced matchings*, Discrete Appl. Math., 24 (1989), pp. 97–102.

[2] E. J. Cockayne, O. Fararon, C. Payan, and A. Thomason, *Contributions to the theory of domination, independence and irredundance in graphs*, Discrete Math., 33 (1981), pp. 249–258.

[3] E. J. Cockayne, S. T. Hedetniemi, and D. J. Miller, *Properties of hereditary hypergraphs and middle graphs*, Canad. Math. Bull., 21 (1978), pp. 461–468.

[4] A. M. Farley and A. Proskurowski, *Computing the maximum order of an open irredundant set in a tree*, Congr. Numer., 41 (1984), pp. 219–228.

[5] A. M. Farley and N. Schacham, *Senders in broadcast networks: Open irredundancy in graphs*, Congr. Numer., 38 (1983), pp. 47–57.

[6] J. F. Fink and M. S. Jacobson, *n-domination in graphs*, in Graph Theory with Applications to Algorithms and Computer Science, John Wiley, New York, 1985, pp. 283–300.

[7] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of* NP-*completeness*, W. H. Freeman, San Francisco, CA, 1979.

[8] M. C. Golumbic and R. Laskar, *Irredundancy in circular arc graphs*, Discrete Appl. Math., 44 (1993), pp. 79–90.

[9] S. E. Goodman, S. Hedetniemi, and R. E. Tarjan, *B-matchings in trees*, SIAM J. Comput., 5 (1976), pp. 104–107.

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

[11] S. M. Hedetniemi, S. T. Hedetniemi, and D. P. Jacobs, *Private domination: Theory and algorithms*, Congr. Numer., 79 (1990), pp. 147–157.

[12] S. T .Hedetniemi, R. C. Laskar, and J. Pfaff, *Irredundance in graphs: A survey*, Congr. Numer., 48 (1985), pp. 183–193.

[13] M. S. Jacobson and K. Peters, *Chordal graphs and upper irredundance, upper domination and independence*, Discrete Math., 86 (1990), pp. 59–69.

[14] ———, *A note on graphs which have upper irredundance equal to independence*, Discrete Appl. Math., 44 (1993), pp. 91–99.

[15] D. Johnson, *The* NP-*completeness column: An ongoing guide*, J. Algorithms, 6 (1985), pp. 434–451.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 48-56, February 1994

BELL INEQUALITIES, GROTHENDIECK'S CONSTANT, AND ROOT TWO*

P. C. FISHBURN† AND J. A. REEDS†

Abstract. B. S. Tsirelson showed that comparisons between probabilities in "classical" physics and probabilities in quantum mechanics yield discrepancy measures K_n for finite $n \times n$ real matrices that approach Grothendieck's constant K_G as n gets large. It is known that $K_2 = K_3 = \sqrt{2}$ and that $K_G \geqq \pi/2 = 1.57\cdots$, but examples of $n \times n$ matrices for specified n that demonstrate $K_n > \sqrt{2}$ have eluded researchers. A series of elementary examples are provided, which yield lower bounds on $K_{k(k-1)}$ that approach $3/2$ as k gets large. A uniform change along the main diagonal of our basic example shows that $K_{20} \geqq 10/7 = 1.42\cdots$.

Key words. Bell inequalities, Grothendieck's constant, finite matrices with large discrepancies, correlation

AMS subject classifications. 05A20, 05B20, 60A05

1. Introduction. Grothendieck's constant K_G, which figures prominently in the theory of linear operators on Banach spaces [11], can be defined as the smallest number such that, for all integers $n \geqq 2$, all $n \times n$ real matrices $[a_{ij}]$, and all $s_1, \ldots, s_n, t_1, \ldots, t_n$ in $\mathbb{R}$ for which

$$\left|\sum_{i,j} a_{ij} s_i t_j\right| \leqq \max_i |s_i| \max_j |t_j|,$$

it is true that

$$\left|\sum_{i,j} a_{ij} \langle x_i, y_j \rangle\right| \leqq K_G \max_i \|x_i\| \max_j \|y_j\|$$

whenever $x_1, \ldots, x_n, y_1, \ldots, y_n$ are vectors in a real Hilbert space. As usual, $\langle x, y \rangle$ is the inner product of x and y, and $\|x\| = \langle x, x \rangle^{1/2}$. The exact value of K_G is unknown. A lower bound of $\pi/2 = 1.5707\cdots$ was established by Grothendieck [6]. We note later that $K_G > \pi/2$. The best upper bound known, due to Krivine [10], is $\pi[2 \log(1 + \sqrt{2})]^{-1} = 1.7822\cdots$.

Our purpose here is to consider Grothendieck-like constants for finite analogues of Grothendieck inequalities that arise from Tsirelson's [1], [13] penetrating comparison between the Bell inequalities [2], [4], [9] for probabilities in "classical" locally deterministic theory with hidden parameters and probabilities that arise in quantum theory [9], [14]. We summarize Tsirelson's connection of this comparison to Grothendieck's constant [13] and then describe our approach and results.

Fix integers $m, n \geqq 2$. For real random variables $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ on a probability space, we define an $m \times n$ matrix $[c_{ij}]$ by $c_{ij} = \mathrm{E}(X_i Y_j)$. Let $C(m, n)$ be the set of all such $[c_{ij}]$ for which

$$\Pr(|X_i| \leqq 1) = \Pr(|Y_j| \leqq 1) = 1 \quad \text{for all } i \text{ and } j$$

and let $Q(m, n)$ be the set of all $[c_{ij}]$ for which

$$\mathrm{E}(X_i^2) \leqq 1 \quad \text{and} \quad \mathrm{E}(Y_j^2) \leqq 1 \qquad \text{for all } i \text{ and } j.$$

* Received by the editors September 16, 1991; accepted for publication (in revised form) January 15, 1993.

† AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

Each of $C(m, n)$ and $Q(m, n)$ is a centrally symmetric convex set in the space of all $m \times n$ real matrices. The extreme points of $C(m, n)$ are rank 1 matrices $[c_{ij}] = [a_i b_j]$ with all $a_i, b_j \in \{1, -1\}$. Examples for $C(2, 2)$ are

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

An important nonextreme point in $C(2, 2)$, obtained by the equally-weighted convex combination of these four extreme points, is

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{1}{2} B_0, \quad \text{where } B_0 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The extreme points of $Q(m, n)$ are more complex, and we refer to §3 of [13] for discussion. An important member of $Q(2, 2)$ is $B_0/\sqrt{2}$. The corresponding Bell's inequality and its quantum analogue [1], [13] for B_0 are, respectively,

$$c_{11} + c_{12} + c_{21} - c_{22} \leqq 2 \quad \text{for } [c_{ij}] \in C(2, 2),$$

$$q_{11} + q_{12} + q_{21} - q_{22} \leqq 2\sqrt{2} \quad \text{for } [q_{ij}] \in Q(2, 2).$$

These identify supporting hyperplanes of $C(2, 2)$ and $Q(2, 2)$, and the ratio of their right sides, i.e., $\sqrt{2}$, measures the discrepancy between $C(2, 2)$ and $Q(2, 2)$ *in the direction* B_0. By finding $K_n > \sqrt{2}$ for larger n, we are implicitly finding inequalities similar to Bell's that show a more striking contrast between quantum and classical correlation matrices.

Tsirelson refers to members of $Q(m, n)$ as quantum realizable correlation matrices; members of $C(m, n)$ are "classical" correlation matrices. In all cases, $C(m, n)$ is properly included in $Q(m, n)$, and $K(m, n)$, defined as the smallest number for which

$$Q(m, n) \subseteq K(m, n) C(m, n)$$

reflects the deficiency of the "classical" approach in the quantum setting. It is known that $K(2, 2) = K(m, 2) = K(2, n) = \sqrt{2}$, as illustrated by $B_0/\sqrt{2}$ and $B_0/2$; that $K(m, 3) = K(3, n) = \sqrt{2}$ for all $m, n \geqq 2$, which Kemperman [8] observes as a consequence of a result of Garg [5]; that $K(m, n)$ is nondecreasing in m and n; and that

$$\lim_{m,n \to \infty} K(m, n) = K_G,$$

which completes the connection to Grothendieck's constant. Henceforth, let $K_n = K(n, n)$.

Tsirelson [13] is careful to note that K_n differs from $K_G(n)$, as defined by Krivine [10], even though $K_G(2) = \sqrt{2}$. Krivine's $K_G(n)$ is the smallest constant in place of K_G in the opening paragraph when all x_i and y_j are restricted to an n-dimensional real Hilbert space. Connections between $K_G(n)$ and the $K(m, n)$ are explained in [13], and we do not pursue them here.

We focus instead on $K(m, n)$ and K_n. Largely as a matter of computational convenience, we adopt a dual approach to that outlined above, which characterizes $K(m, n)$ as the smallest number such that, for all $m \times n$ real matrices $B = [b_{ij}]$ and all unit vectors $\eta_1, \ldots, \eta_n$ in a real Hilbert space,

$$\sum_{i=1}^{m} \left\| \sum_{j=1}^{n} b_{ij}\eta_j \right\| \leqq K(m, n) \max_{\delta_i, \varepsilon_j \in \{1, -1\}} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta_i \varepsilon_j b_{ij}.$$

Since no generality is lost by taking the η_j in $\mathbb{R}^{m+n}$, assume this henceforth. As noted above,

$$K_2 = K_3 = \sqrt{2} \quad \text{and} \quad K_n \to K_G;$$
$$1.5707\cdots \leqq K_G \leqq 1.7822\cdots.$$

Tsirelson and Kemperman have raised the problem of identifying specific B matrices of small dimensions, which imply that

$$K(m, n) > \sqrt{2} = 1.41421\cdots.$$

Since K_n increases beyond 1.57, this might seem easy, but it has been surprisingly difficult. Generic examples that force $K_n > \sqrt{2}$ can be described, but it has not been known how large n must be for this to occur. Extensive calculation and Monte Carlo search for small $m \times n$ matrices by several people failed to uncover a single example for which

$$\max_{\eta_1,\ldots,\eta_n} \sum_i \left\| \sum_j b_{ij}\eta_j \right\| \Big/ \max_{\delta,\varepsilon} \sum_{i,j} \delta_i\varepsilon_j b_{ij} > \sqrt{2}.$$

The first successful example of modest size was discovered by Reeds and N. J. A. Sloane in April 1990. They used a 120×120 B matrix formed from inner products of half the 240 eight-dimensional minimal vectors in the lattice E_8 [3] to obtain $K_{120} \geqq 45/31 = 1.4516\cdots$. We subsequently discovered that similar results are obtained when B is formed from inner products of vectors in related root systems [7, pp. 42, 64, 65].

In particular, we consider B generated by inner products of vectors in D_l, namely, $(0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$ and $(0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$. Let $F_k = \{f_1, f_2, \ldots, f_{k(k-1)}\}$ for $k \geqq 2$ be the set of all k-dimensional vectors with two nonzero components, either 1 and 1, or 1 and -1 in that order, and define the $k(k-1) \times k(k-1)$ B matrix $[f_{ij}^0]$ by $f_{ij}^0 = \langle f_i, f_j \rangle$. We can view F_k as the union of $\binom{k}{2}$ copies of B_0 embedded in

$$\begin{bmatrix} 0\cdots 0 \\ 0\cdots 0 \end{bmatrix}_k.$$

With $m = n = k(k-1)$, we also let

$$N^{(k)} = \max_{\eta_1,\ldots,\eta_n} \sum_{i=1}^{n} \left\| \sum_{j=1}^{n} f_{ij}^0 \eta_j \right\|,$$

$$D^{(k)} = \max_{\delta_i,\varepsilon_j \in \{1,-1\}} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i\varepsilon_j f_{ij}^0.$$

The following result is proved in the next section.

THEOREM 1. $N^{(k)}/D^{(k)} = (3k-3)/(2k-1)$ *for* $k \geqq 2$.

It follows immediately from our dual characterization of K_n that

$$K_{k(k-1)} \geqq (3k-3)/(2k-1) \quad \text{for } k \geqq 2.$$

The smallest k for which this lower bound ratio exceeds $\sqrt{2}$ is $k = 10$, where $27/19 = 1.421\cdots$, so $K_{90} > \sqrt{2}$. Moreover, we now have an explicit finite example of a lower bound on K_n that becomes arbitrarily close to 1.5.

In addition, we have discovered that modification of the main diagonal of $[f_{ij}^0]$ gives a smaller B matrix whose N/D ratio exceeds $\sqrt{2}$. As defined above, $f_{ii}^0 = 2$. By replacing f_{ii}^0 with $f'_{ii} = \frac{2}{3}$ for all i at $k = 5$, the correspondingly modified $N^{(5)}/D^{(5)}$

increases from $12/9$ to $10/7 = 1.428\cdots$ and shows that $K_{20} > \sqrt{2}$. The analysis for this case is included in §3. Reeds [12] subsequently used the diagonal modification idea to prove that $K_G > 1.67$.

As things stand, $n = 20$ is the smallest n for which it is known that $K_n > \sqrt{2}$. The specific B matrix of the preceding paragraph that accomplishes this is shown in Fig. 1 based on $f_1 = 11000, f_2 = 10100, f_3 = 01100, \ldots, f_{10} = 00011, f_{11} = 1-1000, \ldots$, and $f_{20} = 0001-1$.

2. Proof of Theorem 1. We begin with two general lemmas for B matrices determined by inner products and then use these to evaluate $D^{(k)}$. This is followed by our analysis for $N^{(k)}$.

LEMMA 1. *Suppose that $v_i \in \mathbb{R}^M$ are nonzero vectors for $i = 1, \ldots, m$. Let $[b_{ij}]$ be the $m \times m$ matrix with $b_{ij} = \langle v_i, v_j \rangle$ and define D by*

$$D = \max_{\delta_i, \varepsilon_j \in \{1,-1\}} \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_i \varepsilon_j b_{ij}.$$

Let

$$D^* = \max_{\varepsilon_j} \sum_{a=1}^{M} \left(\sum_{j=1}^{m} \varepsilon_j v_{ja} \right)^2 .$$

Then $D^* = D$.

Proof. Clearly, $D^* \leqq D$. The reverse inequality follows from the Cauchy-Schwarz inequality after b_{ij} in D is replaced by $\langle v_i, v_j \rangle$, and the terms in k and in j are grouped separately. □

LEMMA 2. *A maximizing $(\varepsilon_1, \ldots, \varepsilon_m)$ for D under the hypotheses of Lemma 1 satisfies*

$$\varepsilon_i = \operatorname{sgn} \langle v_i, \alpha \rangle \quad \textit{for some } \alpha \in \mathbb{R}^M, \ i = 1, \ldots, m.$$

2/3	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	0	0	0
1	2/3	1	1	1	0	0	1	1	0	1	0	1	1	−1	0	0	1	1	0
1	1	2/3	1	0	1	0	1	0	1	1	1	0	1	0	−1	0	−1	0	1
1	1	1	2/3	0	0	1	0	1	1	1	1	1	0	0	0	−1	0	−1	−1
1	1	0	0	2/3	1	1	1	1	0	−1	−1	0	0	0	1	1	1	1	0
1	0	1	0	1	2/3	1	1	0	1	−1	0	−1	0	1	0	1	−1	0	1
1	0	0	1	1	1	2/3	0	1	1	−1	0	0	−1	1	1	0	0	−1	−1
0	1	1	0	1	1	0	2/3	1	1	0	−1	−1	0	−1	−1	0	0	1	1
0	1	0	1	1	0	1	1	2/3	1	0	−1	0	−1	−1	0	−1	1	0	−1
0	0	1	1	0	1	1	1	1	2/3	0	0	−1	−1	0	−1	−1	−1	−1	0
0	1	1	1	−1	−1	−1	0	0	0	2/3	1	1	1	−1	−1	−1	0	0	0
1	0	1	1	−1	0	0	−1	−1	0	1	2/3	1	1	1	0	0	−1	−1	0
1	1	0	1	0	−1	0	−1	0	−1	1	1	2/3	1	0	1	0	1	0	−1
1	1	1	0	0	0	−1	0	−1	−1	1	1	1	2/3	0	0	1	0	1	1
1	−1	0	0	0	1	1	−1	−1	0	−1	1	0	0	2/3	1	1	−1	−1	0
1	0	−1	0	1	0	1	−1	0	−1	−1	0	1	0	1	2/3	1	1	0	−1
1	0	0	−1	1	1	0	0	−1	−1	−1	0	0	1	1	1	2/3	0	1	1
0	1	−1	0	1	−1	0	0	1	−1	0	−1	1	0	−1	1	0	2/3	1	−1
0	1	0	−1	1	0	−1	1	0	−1	0	−1	0	1	−1	0	1	1	2/3	1
0	0	1	−1	0	1	−1	1	−1	0	0	0	−1	1	0	−1	1	−1	1	2/3

FIG. 1

Proof. Suppose that max for D occurs at $\varepsilon_1^*, \ldots, \varepsilon_m^*$. Let $\alpha_a = \sum_j \varepsilon_j^* v_{ja}$ and $\alpha = (\alpha_1, \ldots, \alpha_M)$. Multiplying each ε_i^* separately by -1 gives $D \geqq \langle \alpha, \alpha \rangle - 4\varepsilon_i^* \langle v_i, \alpha \rangle + 4\langle v_i, v_i \rangle$ for $i = 1, \ldots, m$. Hence

$$\varepsilon_i^* \langle v_i, \alpha \rangle \geqq \langle v_i, v_i \rangle > 0 \quad \text{for } i = 1, \ldots, m. \qquad \square$$

The hyperplanes $\{\alpha \in \mathbb{R}^M : \langle v_i, \alpha \rangle = 0\}$ for $i = 1, \ldots, m$ jointly partition $\mathbb{R}^M$ into subspaces, which we refer to as *cells*, in which no $\langle v_i, \alpha \rangle = 0$. Each cell is the intersection of m open half-spaces. To find a maximizing ε pattern for D, it suffices by Lemma 2 to check one α per cell. Since there may be as many as 2^m cells, this could be burdensome. It can simplify greatly, however, under symmetry.

A case in point is provided by F_k and $D^{(k)}$. To induce full symmetry in the generating basis, let

$$-F_k = \{-f_j : f_j \in F_k\};$$

so $F_k \cup (-F_k)$ is the set of all k-tuples with two nonzero components in $\{1, -1\}$. Let $n = k(k-1)$ and let $[f]^+$ be the $(2n) \times (2n)$ inner product matrix for $F_k \cup (-F_k)$, shown below:

$$\begin{array}{c|cc} & F_k & -F_k \\ \hline F_k & [f_{ij}^0] & -[f_{ij}^0] \\ -F_k & -[f_{ij}^0] & [f_{ij}^0] \end{array}.$$

Also, let D^+ be the maximum value of $\sum_a (\sum_j \varepsilon_j b_{ja})^2$ for $[f]^+$. Suppose that $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are maximizers for $D^{(k)}$. Then maximizers for D^+ are clearly $\varepsilon_1^*, \ldots, \varepsilon_n^*, -\varepsilon_1^*, \ldots, -\varepsilon_n^*$, and

$$D^+ = 4D^{(k)}.$$

By symmetry, $F_k \cup (-F_k)$ is invariant under permutations of coordinates in the permutation group $\mathcal{S}_k$ and invariant to inversions of coordinates. It follows that the cells for α described after the proof of Lemma 2 are isomorphic with respect to the group $\mathcal{S}_k \times \{1, -1\}^k$. It therefore suffices to take $\alpha_1 > \alpha_2 > \cdots > \alpha_k > 0$ to determine a maximizing set of ε_j's for D^+, in which case, since $\langle f_j, \alpha \rangle > 0$ for all $f_j \in F_k$, $\varepsilon_j \equiv 1$ for F_k and $\varepsilon_j \equiv -1$ for $-F_k$. Hence, by $D^+ = 4D^{(k)}$, one set of ε_j maximizes for $D^{(k)}$ is $\varepsilon_1 = \cdots = \varepsilon_n = 1$.

With $\varepsilon_1 = \cdots = \varepsilon_n = 1$ and F_k arranged in the manner used for Fig. 1, we find by Lemma 1 that

$$\begin{aligned} D^{(k)} &= \sum_{a=1}^{k} \left(\sum_{j=1}^{n} f_{ja} \right)^2 = \sum_a [2(k-1) - 2(a-1)]^2 \\ &= 2k(k-1)(2k-1)/3. \end{aligned}$$

We now consider

$$N^{(k)} = \max_{\eta_1, \ldots, \eta_n} \sum_{i=1}^{n} \left\| \sum_{j=1}^{n} f_{ij}^0 \eta_j \right\|$$

with each η_j a unit vector in $\mathbb{R}^{2n}$. It will be shown that one set of maximizing η_j is given by

$$\eta_j^* = f_j / \sqrt{2}, \qquad j = 1, \ldots, n,$$

which use only the first k coordinates. For later use, we note that the value of

$$\sum_i \left\| \sum_j f_{ij}^0 \eta_j^* \right\|$$

is $2k(k-1)^2$. Observe that

$$\sqrt{2} \sum_{j=1}^{n} f_{ij}^0 \eta_j^* = \sum_j \langle f_i, f_j \rangle (f_{j1}, \ldots, f_{jk})$$

$$= \sum_a f_{ia} \left(\sum_j f_{j1} f_{ja}, \ldots, \sum_j f_{jk} f_{ja} \right).$$

Since $k-1$ $f_j = (0 \cdots 1\ 0\ 1 \cdots 0)$ have 1 in position a, and $k-1$ $f_j = (0 \cdots 1\ 0\ {-1} \cdots 0)$ have 1 or -1 in position a, $\Sigma_j f_{ja}^2 = 2(k-1)$. Moreover, if $b \neq a$, then $\Sigma_j f_{jb} f_{ja} = 1 - 1 = 0$; so

$$\sqrt{2} \sum_j f_{ij}^0 \eta_j^* = \sum_a f_{ia} (0, \ldots, 0, 2(k-1)_a, 0, \ldots, 0).$$

Because $f_{ia}^2 = 1$ for two a, but $f_{ia}^2 = 0$ otherwise,

$$\left\| \sum_j f_{ij}^0 \eta_j^* \right\| = [4(k-1)^2 + 4(k-1)^2]^{1/2} / \sqrt{2} = 2(k-1);$$

so $\Sigma_i \| \Sigma_j f_{ij}^0 \eta_j^* \| = 2k(k-1)^2$.

With respect to the maximization of $\Sigma_i \| \Sigma_j f_{ij}^0 \eta_j \|$ over unit vectors η_j in $\mathbb{R}^T$, let

$$\gamma_i = \left\| \sum_j f_{ij}^0 \eta_j \right\| = \left[\sum_{i=1}^{T} \left(\sum_{j=1}^{n} \langle f_i, f_j \rangle \eta_{jt} \right)^2 \right]^{1/2}.$$

Suppose that max $\Sigma \gamma_i^2 = M$. Then $\Sigma \gamma_i$ is maximum when $\gamma_1 = \cdots = \gamma_n = (M/n)^{1/2}$. So we consider maximization of $\Sigma \gamma_i^2$. Expansions of the squared terms give

$$\Sigma \gamma_i^2 = \sum_{t=1}^{T} \sum_{j,p=1}^{n} \sum_{a,b=1}^{k} f_{ja} f_{pb} \eta_{jt} \eta_{pt} \sum_{i=1}^{n} f_{ia} f_{ib}.$$

Since $\Sigma_i f_{ia} f_{ib} = 2(k-1)$ if $a = b$, and is 0 otherwise,

$$\Sigma \gamma_i^2 = 2(k-1) \sum_{j,p} \langle f_j, f_p \rangle \langle \eta_j, \eta_p \rangle.$$

Because the desired result is obvious when $k = 2$, assume that $k \geqq 3$. Separating out the $j = p$ diagonal in the preceding sum, we have

$$\sum_{j,p} \langle f_j, f_p \rangle \langle \eta_j, \eta_p \rangle = 2n + 2 \sum_{j<p} \langle f_j, f_p \rangle \langle \eta_j, \eta_p \rangle.$$

Choose any three columns from the $n \times k$ matrix for F_k. These columns correspond to six f_i that contribute nonzero values of $\langle f_j, f_p \rangle$ for $j < p$, e.g.,

$$\begin{array}{lrrr}
f_1: & 1 & 1 & 0, \\
f_2: & 1 & 0 & 1, \\
f_3: & 0 & 1 & 1, \\
f_4: & 1 & -1 & 0, \\
f_5: & 1 & 0 & -1, \\
f_6: & 0 & 1 & -1,
\end{array}$$

with the f's subscripted for convenience. It is easily seen that each nonzero $\langle f_j, f_p\rangle$ for $j < p$ arises exactly once in the $\binom{k}{3}$ choices of three columns. The contribution to $\Sigma_{j<p} f^0_{jp}\langle \eta_j, \eta_p\rangle$ for the three illustrated columns is, with $\eta^0_{jp} = \langle \eta_j, \eta_p\rangle$,

$$(\eta^0_{12} + \eta^0_{16} - \eta^0_{26}) + (\eta^0_{13} + \eta^0_{15} - \eta^0_{35}) + (\eta^0_{23} + \eta^0_{24} - \eta^0_{34}) + (\eta^0_{45} + \eta^0_{56} - \eta^0_{46}).$$

Each of the four terms in parentheses has similar structure, and it is routine but tedious to check that the maximum of each is 3/2. For example, max $(\eta^0_{12} + \eta^0_{16} - \eta^0_{26}) = 3/2$ at

$$\begin{array}{llll} \eta_1: & 1/\sqrt{2} & 1/\sqrt{2} & 0, \\ \eta_2: & 1/\sqrt{2} & 0 & 1/\sqrt{2}, \\ \eta_6: & 0 & 1/\sqrt{2} & -1/\sqrt{2}. \end{array}$$

It follows that

$$\max \Sigma\gamma_i^2 \leqq 2(k-1)\left[2n + 2\binom{k}{3}4(3/2)\right] = 4n(k-1)^2$$

As remarked earlier, $\Sigma\gamma_i \leqq n(M/n)^{1/2}$; so

$$\max \Sigma\gamma_i \leqq n[4n(k-1)^2/n]^{1/2} = 2k(k-1)^2.$$

Since the η_j^* attain $2k(k-1)^2$ for $\Sigma\gamma_i$, as shown earlier,

$$N^{(k)} = 2k(k-1)^2.$$

Therefore $N^{(k)}/D^{(k)} = 2k(k-1)^2/[2k(k-1)(2k-1)/3] = (3k-3)/(2k-1)$.

3. Diagonal modification. Given $\lambda \in [0, 2]$, we consider the effects on the results of the preceding section when λ is subtracted from each element on the main diagonal of $[f^0_{ij}]$. With $F_k = \{f_1, \ldots, f_n\}$, $n = k(k-1)$, let $[f^\lambda_{ij}]$ be the $n \times n$ matrix for which

$$f^\lambda_{ij} = \langle f_i, f_j\rangle, \qquad i \neq j,$$
$$f^\lambda_{ii} = \langle f_i, f_i\rangle - \lambda = 2 - \lambda.$$

Also, define

$$N^{(k)}_\lambda = \max_{\eta_1,\ldots,\eta_n} \sum_{i=1}^n \left\| \sum_{j=1}^n f^\lambda_{ij}\eta_j \right\|,$$

$$D^{(k)}_\lambda = \max_{\delta_i,\varepsilon_j \in \{1,-1\}} \sum_{i=1}^n \sum_{j=1}^n \delta_i\varepsilon_j f^\lambda_{ij},$$

where the η_j are unit vectors. Hence $K_n \geqq N^{(k)}_\lambda/D^{(k)}_\lambda$ for all $k \geqq 2$.

The effect of λ on $N^{(k)}$ is straightforward. With $\eta_j^* = f_j/\sqrt{2}$ as in the preceding section, calculation gives

$$\sum_{i=1}^n \left\| \sum_{j=1}^n f^\lambda_{ij}\eta_j^* \right\| = n[2(k-1) - \lambda].$$

The approach described in the final paragraph of the preceding section yields

$$\begin{aligned}\Sigma\gamma_i^2 &= [2(k-1)-2\lambda]\sum_{j,p}\langle f_j,f_p\rangle\langle\eta_j,\eta_p\rangle+\lambda^2 n\\ &\leqq [2(k-1)-2\lambda]\left[2n+2\binom{k}{3}4(3/2)\right]+\lambda^2 n\\ &= n[2(k-1)-\lambda]^2,\end{aligned}$$

and therefore $\Sigma\gamma_i \leqq n\{n[2(k-1)-\lambda]^2/n\}^{1/2} = n[2(k-1)-\lambda]$. Hence

$$N_\lambda^{(k)} = n[2(k-1)-\lambda] = N^{(k)}-\lambda n.$$

The situation with $D_\lambda^{(k)}$ is more delicate. Let $\delta(a) = \sum_i \delta_i f_{ia}$ and $\varepsilon(a) = \sum_i \varepsilon_i f_{ia}$, so that

$$D_\lambda^{(k)} = \max_{\delta,\varepsilon}\left[\sum_{a=1}^{k}\delta(a)\varepsilon(a)-\lambda\sum_{i=1}^{n}\delta_i\varepsilon_i\right].$$

We proved earlier that $\delta_i = \varepsilon_i = 1$ for all i give $D_0^{(k)} = D^{(k)} = 2n(2k-1)/3$, and it is easily seen that

$$D_\lambda^{(k)} = D^{(k)}-\lambda n \quad \text{for small } \lambda > 0.$$

However, as λ increases, $D_\lambda^{(k)}$ changes form because the maximizing δ_i and ε_i do not remain uniformly 1.

To consider other sign patterns for δ and ε, let

$$A_s = \{(\delta,\varepsilon) : |\{i : \delta_i\varepsilon_i = -1\}| = s\}.$$

Then $\Sigma\delta_i\varepsilon_i = n-2s$ for each $(\delta,\varepsilon)\in A_s$. Also, let

$$D(s) = \max_{(\delta,\varepsilon)\in A_s}\sum_{a=1}^{k}\delta(a)\varepsilon(a),$$

where supeiscript (k) on D is omitted for convenience. Then $D(s)-\lambda(n-2s)$ is the maximum value of

$$\left[\sum_{a=1}^{k}\delta(a)\varepsilon(a)-\lambda\sum_{i=1}^{n}\delta_i\varepsilon_i\right]$$

when (δ,ε) is in A_s, and

$$D_\lambda^{(k)} = \max\{D(0)-\lambda n, D(1)-\lambda(n-2), D(2)-\lambda(n-4), \ldots\}.$$

Here $D(0) = D^{(k)}$.

It is easily seen that $D(s)$-maximizing sign changes for the first few s are as follows:

		δ_i	ε_i	
$s = 1$:	$1\ -1\ 0\cdots 0$	1	-1	$D(1) = D(0) - 4,$
$s = 2$:	$1\ -1\ 0\ 0\cdots 0$	1	-1	$D(2) = D(0) - 8,$
	$1\ 0\ -1\ 0\cdots 0$	-1	1	
$s = 3$:	$1\ -1\ 0\ 0\cdots 0$	1	-1	
	$1\ 0\ -1\ 0\cdots 0$	-1	1	$D(3) = D(0) - 8.$
	$0\ 1\ -1\ 0\cdots 0$	1	-1	

The most salient feature of these changes is that $D(3) = D(2)$, i.e., that the decrease of $D(0)$ by 8 at $s = 2$ does not change when we go to $s = 3$. As a result, $D(3) - \lambda(n-6) > D(2) - \lambda(n-4)$, so that $D_\lambda^{(k)}$ never obtains at $s = 2$.

By comparing $D(0) - \lambda n$ at $s = 0$ and $D(3) - \lambda(n-6) = D(0) - 8 - \lambda(n-6)$ at $s = 3$, we see that

$$\begin{aligned}\max\{D(0) - \lambda n, D(3) - \lambda(n-6)\} &= D(0) - \lambda n \quad \text{if } \lambda < 4/3,\\ &= D(3) - \lambda(n-6) \quad \text{if } \lambda > 4/3.\end{aligned}$$

The break-even point of $\lambda = 4/3$ corresponds to $2/3$ on the main diagonal of $[f_{ij}^\lambda]$, as in Fig. 1. In addition, for $\lambda \in [0, 2]$, we have checked for $k \in \{3, 4, 5\}$ that the preceding max is in fact $D_\lambda^{(k)}$ and that the ratio $N_\lambda^{(k)}/D_\lambda^{(k)} = [N^{(k)} - \lambda n]/\max\{D(0) - \lambda n, D(3) - \lambda(n-6)\}$ is maximized at $\lambda = 4/3$ where the denominator changes slope. That is,

$$\max_{0 \leqq \lambda \leqq 2} N_\lambda^{(k)}/D_\lambda^{(k)} = N_{4/3}^{(k)}/D_{4/3}^{(k)} = \frac{3k-5}{2k-3} \quad \text{for } k \in \{3, 4, 5\}.$$

We suspect that the same result holds for $k > 5$, but have not proved this.

The ratio $(3k-5)/(2k-3)$ is less than $\sqrt{2}$ for $k < 5$ but equals $10/7 = 1.428\cdots$ at $k = 5$. Hence $K_{5(4)} = K_{20} \geqq 10/7$, and $n = 20$ is presently the smallest known n for which $K_n > \sqrt{2}$.

Acknowledgments. We are indebted to several people for collaboration on the research reported here, including N. J. A. Sloane, L. A. Shepp, C. L. Mallows, J. H. B. Kemperman, B. S. Tsirelson, and L. A. Khalfin. Special thanks go to L. A. Shepp for his English translation of [13].

REFERENCES

[1] B. S. CIREL'SON (TSIRELSON), *Quantum generalizations of Bell's inequality*, Lett. Math. Phys., 4 (1980), pp. 93–110.

[2] J. F. CLAUSER AND A. SHIMONY, *Bell's theorem: Experimental tests and implications*, Rep. Prog. Phys., 41 (1978), pp. 1881–1927.

[3] J. H. CONWAY AND N. J. A. SLOANE, *Sphere Packings, Lattices and Groups*, Springer-Verlag, New York, 1988.

[4] M. FROISSART, *Constructive generalization of Bell's inequalities*, Il Nuovo Cimento, 64B (1981), pp. 241–251.

[5] A. GARG, *Detector error and Einstein-Podolsky-Rosen correlations*, Phys. Rev. D, 28 (1983), pp. 785–790.

[6] A. GROTHENDIECK, *Reśumé de la théorie métrique des produits tensoriels topologiques*, Bol. Soc. Mat. São-Paulo, 8 (1956), pp. 1–79.

[7] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, 2nd printing, revised, Springer-Verlag, New York, 1972.

[8] J. H. B. KEMPERMAN, personal communication, 1990.

[9] L. A. KHALFIN AND B. S. TSIRELSON, *Quantum/classical correspondence in the light of Bell's inequalities*, Found. Phys., 22 (1992), pp. 879–948.

[10] J. L. KRIVINE, *Constantes de Grothendieck et fonctions de type positif sur les sphères*, Adv. Math., 31 (1979), pp. 16–30.

[11] G. PISIER, *Factorization of Linear Operators and Geometry of Banach Spaces*, American Mathematical Society, Providence, RI, 1986.

[12] J. A. REEDS, *A New Lower Bound on the Real Grothendieck Constant*, preprint, AT&T Bell Laboratories, Murray Hill, NJ, 1992.

[13] B. S. TSIRELSON, *Quantum analogues of the Bell inequalities: The case of 2 spatially separated regions*, in Problems of the Theory of Probability Distributions IX, Math. Inst. Steklov (LOMI), 142 (1985), pp. 174–194. (In Russian.) (English transl., Quantum analogues of the Bell inequalities: The case of two spatially separated domains, J. Soviet Math., 36 (1987), pp. 557–570.)

[14] E. WIGNER, *On hidden parameters and quantum mechanical probabilities*, in Studies of Symmetry, pp. 294–302, 1971. (In Russian.)

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 57–71, February 1994

EXTREMAL PROBLEMS IN THE CONSTRUCTION OF DISTRIBUTED LOOP NETWORKS*

D. FRANK HSU† AND XING-DE JIA‡

Abstract. Let $G(N, A)$ be the Cayley digraph associated with $\mathbf{Z}/(N)$ and A, where N is a positive integer and A is a subset of $\{1, 2, \ldots, N-1\}$. Let $N(d, k)$ be the maximum N such that the diameter of $G(N, A)$ is less than or equal to d for some $A = \{a_1, a_2, \ldots, a_k\}$ with $1 = a_1 < a_2 < \cdots < a_k$. An exact formula for $N(d, 2)$ is given, and $N(d, k)$ is estimated for $k \geq 3$. These results provide new bounds for minimal diameter in the construction of loop networks. A relation between this problem and the postage stamp problem in additive number theory is established to enhance the study of these problems.

Key words. Cayley digraphs, distributed loop networks, extremal problems, diameter, postage stamp problem, combinatorial optimization

1. Introduction. Let a and b be integers. Let $[a, b]$ denote the set of all integers x such that $a \leq x \leq b$. Similarly, $(a, b]$ and $[a, b)$ are the sets of integers $x \in [a, b]$, excluding a and b, respectively. We denote as $\lfloor x \rfloor$ the greatest integer $\leq x$, and as $\lceil x \rceil$ the least integer $\geq x$. Let N be a positive integer. Let A be a subset of $[1, N-1]$. Let $\mathbf{Z}/(N)$ be the additive group of residue classes modulo N. A *Cayley digraph* $G(N, A)$ associated with $\mathbf{Z}/(N)$ and A has its vertices labeled with elements in $\mathbf{Z}/(N)$, and vertex i is adjacent to vertex j if and only if $j = i + a$ for some $a \in A$.

For a given digraph $G(N, A)$, let $d(G(N, A))$ be its diameter. Let

$$d(N, k) = \min \{d(G(N, A)) | \text{ for all } A \text{ with } |A| = k\}.$$

On the other hand, for any given $d \geq 2$, we define $N(d, k)$ as the maximum N such that there is an

$$A = \{1 = a_1 < a_2 < \cdots < a_k\}$$

of k elements such that $d(G(N, A)) \leq d$. In other words, $N(d, k)$ is the maximal number of vertices so that the outdegree of the digraph is k and the diameter is not greater than d. In this paper, we are most concerned with the following problems:

(a) For any given $k \geq 2$, find $d(N, k)$ in terms of N;

(b) Given $k \geq 2$, find $N(d, k)$ in terms of d.

Cayley digraphs as defined above have recently attracted much attention for their theoretical merits and in applications to distributed loop communications networks and in the construction of massively parallel processors. For further information, refer to Bermond, Comellas, and Hsu [1] and D. V. Chudnovsky, G. V. Chudnovsky, and Denneau [2]. In §2 we study the number $N(d, k)$ in terms of d for given k. In particular, we obtain an exact formula for $N(d, 2)$ and an upper bound and a lower bound for $N(d, 3)$ and we discuss bounds for $N(d, k)$, when $k \geq 4$. In §3 we establish a relation between $N(d, k)$ and the postage stamp problem in additive number theory. In §4, using the results from §§2 and 3, we obtain new bounds for the optimal diameter of distributed loop communication networks, which improve those obtained by Erdös and Hsu [3], Fiol et al. [4], Hwang and Xu [7], and Wong and Coppersmith [13].

2. Bounds for $N(d, k)$. For a given $k \geq 2$, we are interested in computing $N(d, k)$ in terms of d. First, we study the case when $k = 2$. Since the Cayley digraph $G(N, A)$ is

* Received by the editors July 23, 1990; accepted for publication (in revised form) October 20, 1992.

† Department of Computer and Information Science, Fordham University, Bronx, New York 10458-5198.

‡ Department of Mathematics, Southwest Texas State University, San Marcos, Texas 78666.

vertex symmetric, it is sufficient to study the distance from zero to all other vertices in computing the diameter of the digraph. Wong and Coppersmith [13] provided the construction of a pattern for those distances.

In the computation of $d(G(N, A))$, where $A = \{1, s\}$, we proceed to fill the lattice point (x, y) ($x \geq 0$, $y \geq 0$ are integers) of the Euclidean plane with an integer $n (0 \leq n < N)$ if

$$x \cdot 1 + y \cdot s \equiv n \pmod{N}.$$

We start from the origin (0, 0), then the points on the line (1, 0), (0, 1), and then the points on the line (2, 0), (1, 1), (0, 2), and so on. At each point (x, y), if the value n has not appeared so far, we write it down; otherwise, we just leave a blank. The process ends when all values of n in $[0, N - 1]$ have been used. Wong and Coppersmith [13] proved the following lemma.

LEMMA 1 (see Wong and Coppersmith [13]). *The filled pattern is always of the form shown in Fig.* 1, *where* $m \geq 0$, $n \geq 0$, $p > 0$, *and* $q > 0$. *Clearly, the diameter* $d(G(N, A))$ *of* $G(N, A)$, *where* $A = \{1, s\}$, *is equal to* $m + q + \max\{n, p\} - 2$.

We now state and prove Theorem 1.

THEOREM 1. *For any* $d \geq 2$,

$$N(d, 2) = \left\lfloor \frac{d(d+4)}{3} \right\rfloor + 1.$$

Proof. First, we show that

$$N(d, 2) \geq \left\lfloor \frac{d(d+4)}{3} \right\rfloor + 1 = N_1 = N_1(d).$$

It is clear that we need only construct a set $A = \{1, b\}$ so that

$$(1) \qquad dA = \{u \cdot 1 + v \cdot b \mid u \geq 0, v \geq 0, u + v \leq d\} = \mathbf{Z}/(N_1).$$

Suppose that $d = 3t + i$ for some t and $1 \leq i \leq 3$. Then

$$N_1 = \left\lfloor \frac{(3t+i)(3t+i+4)}{3} \right\rfloor + 1 = 3t^2 + (4 + 2i)t + 3i - 1.$$

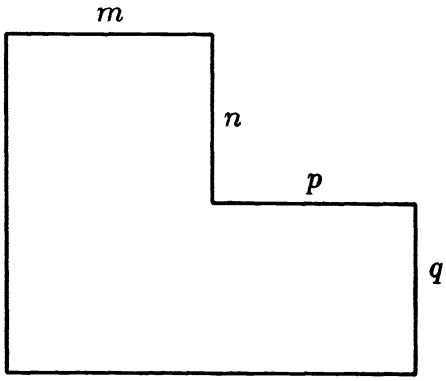


FIG. 1. *The general pattern.*

Let $a = 1$ and $b = d + i = 3t + 2i$. We show that (1) holds. Let $x \in [0, N_1)$. If $x \in [(t+1)b, N_1)$, then

$$\begin{aligned} x - (t+1)b &\leq N_1 - 1 - (t+1)b \\ &= (3t^2 + (4+2i)t + 3i - 1) - 1 - (t+1)(3t+2i) \\ &= t + i - 2. \end{aligned}$$

Noting that

$$(x - (t+1)b) + (t+1) \leq (t+i-2) + (t+1) = 2t + i - 1 < d,$$

we see that $x = (x-(t+1)b)\cdot 1 + (t+1)\cdot b \in dA$. Next, we assume that $x \in [ub, (u+1)b)$ for some $u \in [1, t]$. It is easy to check that the vertex $(u+t+2)b - N_1$ lies between ub and $(u+1)b$.

If $x \in [ub, (u+t+2)b - N_1)$, then

$$\begin{aligned} x - ub &\leq (u+t+2)b - N_1 - 1 - ub = (t+2)b - N_1 - 1 \\ &= (t+2)(3t+2i) - (3t^2 + (4+2i)t + 3i - 1) - 1 \\ &= 2t + i. \end{aligned}$$

Therefore,

$$(x - ub) + u \leq 2t + i + t = 3t + i = d,$$

which implies that $x = (x - ub)\cdot 1 + u\cdot b \in dA$.

If $x \in [(u+t+2)b - N_1, (u+1)b)$, then

$$\begin{aligned} x - ((u+t+2)b - N_1) &\leq (u+1)b - 1 - ((u+t+2)b - N_1) \\ &= N_1 - 1 - (t+1)b \\ &= 3t^2 + (4+2i)t + 3i - 1 - 1 - (t+1)(3t+2i) \\ &= t + i - 2. \end{aligned}$$

Noting that

$$(x - (u+t+2)b - N_1) + (u+t+2) \leq (t+i-2) + (t+t+2) = 3t + i = d,$$

we have

$$x \equiv x + N_1 = (x - (u+t+2)b - N_1)\cdot 1 + (u+t+2)\cdot b \in dA.$$

Therefore, $N(d, 2) \geq N_1$.

We now show that N_1 is indeed also an upper bound for $N(d, 2)$. Suppose that $A = \{1, b\}$ is such that $d(G(N(d, 2), A) \leq d$, i.e.,

$$dA = \{u\cdot 1 + v\cdot b \mid u \geq 0, v \geq 0, u + v \leq d\} = \mathbf{Z}/(N),$$

where $N = N(d, 2)$. We need to show that $N \leq N_1$. We divide the proof into three cases.

Case 1. $d \equiv 0 \pmod 3$. Let $d = 3t$ for some t. In this case, by Lemma 1, the best possible values for m, q, and $\max(n, p)$ are $m = q = t + 1$ and $n = p = t$. Hence

$$N = (t+1)^2 + 2(t+1)t = 3t^2 + 4t + 1 = N_1.$$

Case 2. $d \equiv 1 \pmod 3$. Let $d = 3t + 1$. Under the condition that

$$m + q + \max(n, p) - 2 \leq d = 3t + 1$$

in Lemma 1, the best possible patterns for Fig. 1 with maximum area are (i) $m = q = p = n = t + 1$, and (ii) $m = t + 1$, $q = t + 2$, $p = n = t$. In (ii), we have

$$N = (t + 1)(t + 2) + t(t + 1) + t(t + 2) = 3t^2 + 6t + 2 = N_1.$$

The pattern (i) gives a larger N, but we now show that (i) is not possible. If this pattern is possible, then the number of lattice points in the pattern is $N_0 = 3(t + 1)^2$. The lattice point $(t + 1, t + 1)$ at the corner must represent 0 (mod N_0) (see Fig. 2). Hence

$$(t + 1)\cdot 1 + (t + 1)\cdot b \equiv 0 \pmod{N_0},$$

which implies that

$$1 + b \equiv 0 \pmod{3(t + 1)}.$$

We may assume that $1 + b = 3(t + 1)u$, where $1 \le u < (t + 1)$. We claim that gcd $(u, t + 1) = 1$. Otherwise, let $u = ru'$, $t + 1 = rt'$ for some r, $1 < r < t + 1$. Then

$$t'\cdot 1 + t'\cdot b = t'(1 + b) = t'\cdot 3(t + 1)u = 3(t + 1)^2u' \equiv 0 \pmod{N_0};$$

i.e., the point (t', t') represents 0. This is a contradiction because this point is contained inside the pattern. Therefore, the equation

$$uy \equiv 1 \pmod{(t + 1)}$$

has a solution x_0: $t + 2 \le x_0 \le 2t + 2$. Consider the lattice point $(0, x_0)$ inside the pattern. Noting that

$$\begin{aligned} x_0b &= x_0(1 + b) - x_0 = x_0\cdot 3(t + 1)u - x_0 \\ &\equiv 3(t + 1) - x_0 \pmod{N_0} \end{aligned}$$

and

$$t + 1 \le 3(t + 1) - x_0 \le 2t + 1,$$

we see that $(0, x_0)$ and $(3(t + 1) - x_0, 0)$ in the pattern represent the same value $x_0b \equiv 3(t + 1) - x_0$. This is a contradiction.

Case 3. $d \equiv 2 \pmod 3$. Suppose that $d = 3t + 2$. It follows from Lemma 1 that the best possible pattern is the one with

$$m = n = p = t + 1 \quad \text{and} \quad q = t + 2.$$

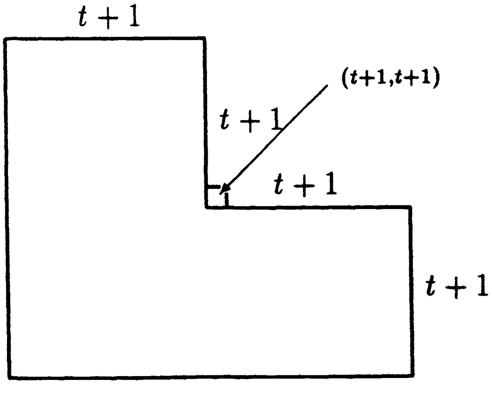


FIG. 2. *The pattern with* $m = q = p = n = t + 1$.

Then

$$N = (t+1)(t+2) + (t+1)^2 + (t+1)(t+2) = 3t^2 + 8t + 5 = N_1.$$

Combining Cases 1–3, we have that $N(d, 2) \geq N_1$. This completes the proof of Theorem 1. □

We now return to the case when $k = 3$. First, we give a lower bound for $N(d, 3)$.

THEOREM 2. *$N(d, 3) \geq \frac{1}{16}d^3 + \frac{3}{8}d^2 + O(d)$ as $d \to \infty$.*

Proof. We construct a set A of three elements so that

$$d(G(N_1, A)) \leq d, \quad \text{where } N_1 \geq \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + O(d).$$

Equivalently, we must construct $A = \{1, b, c\}$ such that

$$dA = \{u \cdot 1 + v \cdot b + w \cdot c \mid u, v, w \geq 0 \text{ and } u + v + w \leq d\} = \mathbf{Z}/(N_1).$$

Let

$$r = \left\lfloor \frac{d}{4} \right\rfloor,$$
$$b = 3d - 8r + 5,$$
$$c = rb + d - 3r + 2,$$
$$N_1 = rc + 2d - 6r + 3.$$

Let $A = \{1, b, c\}$. We show that $dA = \mathbf{Z}/(N_1)$. If $v, w \in [1, r]$, then $r + v + w \leq 3r < d$. Hence, for any $v, w \in [1, r]$, we have

$$vb + wc \in dA, \quad (v + r)b + wc \in dA, \quad \text{and} \quad vb + (w + r)c \in dA.$$

It is clear that

$$\begin{aligned} &[vb + wc, vb + wc + d - v - w] \subseteq dA, \\ (2) \qquad &[(r + v)b + wc, (r + v)b + wc + d - r - v - w] \subseteq dA, \\ &[vb + (r + w)c, vb + (r + w)c + d - v - r - w] \subseteq dA. \end{aligned}$$

The basic idea in the proof is to arrange all these intervals to cover an interval of length N_1 modulo N_1. Since

$$N_1 + d - v - w + 1 \geq N_1 + d - 2r + 1 = rc + b - 1,$$

we see that

$$N_1 + (v - 1)b + wc + d - (v - 1) - w \geq vb + (w + r)c - 1.$$

It therefore follows from (2) that

$$(3) \qquad [N_1 + (v - 1)b + wc, vb + (r + w)c] \subseteq dA.$$

Noting that

$$\begin{aligned} rc + d - v - (w + r) &\geq rc + d - 3r \\ &= (rc + 2d - 6r + 3) - 2d + 6r - 3 + d - 3r \\ &= N_1 - d + 3r - 3 \\ &= N_1 + rb - rb - d + 3r - 3 \\ &= N_1 + rb - c - 1, \end{aligned}$$

we see that

$$vb + (w + r)c + d - v - (w + r) \geq N_1 + (r + v)b + (w - 1)c - 1.$$

This implies that

$$(4) \qquad [vb + (r + w)c, N_1 + (v + r)b + (w - 1)c] \subseteq dA.$$

It follows from (3) and (4) that

$$(5) \qquad [N_1 + (v - 1)b + wc, N_1 + (v + r)b + (w - 1)c] \subseteq dA.$$

Since

$$rb - c + d - v - w + 1 - r \geq rb - c + d - 3r + 1 = -1,$$

we see that

$$N_1 + (r + v)b + (w - 1)c + d - (r + v) - (w - 1) \geq N_1 + vb + wc - 1,$$

which implies that

$$[N_1 + (r + v)b + (w - 1)c, N_1 + vb + wc) \subseteq dA.$$

It therefore follows from (5) that

$$[N_1 + (v - 1)b + wc, N_1 + vb + wc] \subseteq dA.$$

Hence the arbitrariness of v: $1 \leq v \leq r$ implies that

$$(6) \qquad [N_1 + wc, N_1 + rb + wc] \subseteq dA.$$

Noting that, for $d \geq 4$,

$$\begin{aligned} rb + d - w - r + 1 &\geq rb + d - r - r \\ &= c + r - 1 \geq c, \end{aligned}$$

we see that

$$N_1 + rb + wc + d - r - w \geq N_1 + (w + 1)c,$$

which implies that

$$(7) \qquad [N_1 + rb + wc, N_1 + (w + 1)c] \subseteq dA.$$

From (6) and (7), we have

$$[N_1 + wc, N_1 + (w + 1)c) \subseteq dA.$$

Again by the arbitrariness of w: $1 \leq w \leq r$, we see that

$$[N_1 + c, N_1 + (r + 1)c) \subseteq dA.$$

When $d \geq 27$,

$$\begin{aligned} (r + 1)c + d - (r + 1) &= rc + c + d - (r + 1) \\ &= N_1 - 2d + 6r - 3 + c + d - (r + 1) \\ &= N_1 + c - 1 + 5r - d - 3 \\ &> N_1 + c - 1 + 5\left\lfloor \frac{d}{4} \right\rfloor - d - 3 \\ &\geq N_1 + c - 1. \end{aligned}$$

Therefore $[N_1 + c, 2N_1 + c] \subseteq dA$. This means that $dA = \mathbf{Z}/(N_1)$.

Finally, we show that

$$N_1 \geq \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + O(d).$$

In fact,

$$\begin{aligned} N_1 &= rc + 2d - 6r + 3 \\ &= r^2b + rd - 3r^2 + 2r + 2d - 6r + 3 \\ &= r^2(3d - 8r + 2) + r(d - 4) + 2d + 3. \end{aligned}$$

Since $r = \lfloor d/4 \rfloor$, we now divide into four cases according to $d = 4r$, $4r + 1$, $4r + 2$, or $4r + 3$. By substituting r in terms of d in all four cases, we have

(a) $N_1 = \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + d + 3$, if $d = 4r$,
(b) $N_1 = \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + \tfrac{5}{16}d + \tfrac{17}{14}$, if $d = 4r + 1$,
(c) $N_1 = \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 - \tfrac{3}{4}d + \tfrac{13}{2}$, if $d = 4r + 2$,
(d) $N_1 = \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 - \tfrac{25}{16}d + \tfrac{21}{2}$ if $d = 4r + 3$.

In any case, we have $N_1 = \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + O(d)$. Hence

$$N(d, 3) \geq \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + O(d).$$

This completes the proof of Theorem 2. □

Actual calculation gives the value $N(d, 3)$ for each d: $2 \leq d \leq 15$ and the set A for which $N(d, A)$ achieves the maximum $N(d, 3)$. These are listed in Table 1.

Suppose that $A = \{1, b, c\}$ is such that $d(G(N(d, 3), A) \leq d$. We consider the first "quadrant" in the three-dimensional Euclidean space $x \geq 0$, $y \geq 0$, and $z \geq 0$. Let $r \geq 0$ be an integer. Define

$$S_r^{(3)} = \{(x, y, z) \mid x, y, z \geq 0 \text{ integers, and } x + y + z = r\}.$$

In [13] Wong and Coppersmith showed how to visit the lattice points in

$$S_0^{(3)} \cup S_1^{(3)} \cup S_2^{(3)} \cup \cdots.$$

For a given N, $S_r^{(3)}$ is visited in the order of $r = 0, 1, 2, \ldots$. Within each $S_r^{(3)}$, the visitation order $O_r^{(3)}$ is defined recursively as follows:

(i) Let U_i be the set of points in $S_r^{(3)}$ with the third component equal to i. They will be visited in the order $i = 0, 1, 2, \ldots$;

(ii) In U_i, regard the first two components of each point as a point in $S_{r-i}^{(2)}$. Hence U_i can be identified with $S_{r-i}^{(2)}$. Visit U_i in the order $O_{r-i}^{(2)}$;

(iii) $O_r^{(2)}$ is defined as: first, visit $(r, 0)$, then $(r - 1, 1)$, $(r - 2, 2), \ldots, (0, r)$.

TABLE 1
Exact values of N(d, k) with a corresponding A.

d	$N(d, 3)$	A	d	$N(d, 3)$	A
2	9	{1, 3, 4}	9	138	{1, 11, 78}
3	16	{1, 4, 5}	10	176	{1, 17, 56}
4	27	{1, 4, 17}	11	217	{1, 13, 119}
5	40	{1, 6, 15}	12	273	{1, 14, 153}
6	57	{1, 13, 33}	13	340	{1, 90, 191}
7	78	{1, 6, 49}	14	395	{1, 35, 271}
8	111	{1, 31, 69}	15	462	{1, 29, 97}

Now we use the visitation order $O_r^{(3)}$ to fill the lattice point (x, y, z) in $S_r^{(3)}$ of the Euclidean plane with an integer $n \in [0, N)$ if $x \cdot 1 + y \cdot b + z \cdot c \equiv n \pmod{N}$. We start from $S_0^{(3)}$, then the lattice points on the plane $S_1^{(3)}$ using $O_1^{(3)}$, and then the plane $S_2^{(3)}$ using $O_2^{(3)}$, and so on. At each lattice point (x, y, z), if the resulting value n has not appeared before, we write it down. Otherwise, we leave it blank. This process ends when all values of n in $[0, N - 1]$ have been used. We have the following lemma.

LEMMA 2. *The filled pattern is always of the form shown in Fig. 2.*

Proof. We note that, in the above construction, if the lattice point (x_1, y_1, z_1) is blank, then all lattice points (x, y, z) with $x \geq x_1$, $y \geq y_1$, $z \geq z_1$ will also be blank. For example, suppose that the lattice point (x_1, y_1, z_1) represents the value t: $0 \leq t \leq N - 1$, which has appeared before at lattice point (x_0, y_0, z_0) with $x_0 + y_0 + z_0 \leq x_1 + y_1 + z_1$. That is, we have

$$t \equiv x_1 \cdot 1 + y_1 \cdot b + z_1 \cdot c \equiv x_0 \cdot 1 + y_0 \cdot b + z_0 \cdot c \pmod{N}.$$

Then the value $t + 1 = (x_1 + 1) \cdot 1 + y_1 \cdot b + z_1 \cdot c$ at $(x_1 + 1, y_1, z_1)$ must have appeared at $(x_0 + 1, y_0, z_0)$, which is visited before because

$$\begin{aligned} t + 1 &\equiv (x_1 + 1) \cdot 1 + y_1 \cdot b + z_1 \cdot c \\ &\equiv (x_0 + 1) \cdot 1 + y_0 \cdot b + z_0 \cdot c \pmod{N} \end{aligned}$$

and

$$\begin{aligned} (x_0 + 1) + y_0 + z_0 &= (x_0 + y_0 + z_0) + 1 \\ &\leq (x_1 + y_1 + z_1) + 1 = (x_1 + 1) + y_1 + z_1. \end{aligned}$$

Hence $(x_1 + 1, y_1, z_1)$ is blank. Let (x, y, z) be a point that is blank after the visitation. (x, y, z) is called an *x-point* if it is not on the x-axis, and both $(x, y - 1, z)$ and $(x, y, z - 1)$ are in the pattern. (x, y, z) is called an *xy-point* if it is not on the xy-plane, and $(x, y, z - 1)$ is in the pattern. Similarly, we may define y-, z-, yz-, and xz-points.

Let (x, y, z) be an xy-point. Suppose that (x, y, z) represents the same value n as a previously visited point (x_1, y_1, z_1) inside the pattern, i.e.,

$$n = x \cdot 1 + y \cdot b + z \cdot c \equiv x_1 \cdot 1 + y_1 \cdot b + z_1 \cdot c \pmod{N}, \tag{8}$$

where $x_1, y_1, z_1 \geq 0$ and $x_1 + y_1 + z_1 \leq x + y + z$. We now show that (x_1, y_1, z_1) is on the xy-plane, i.e., $z_1 = 0$. Otherwise, we consider the point $(x_1, y_1, z_1 - 1)$, which is a point in the pattern. It follows from (8) that

$$\begin{aligned} x \cdot 1 + y \cdot b + (z - 1) \cdot c &\equiv x_1 \cdot 1 + y_1 \cdot b + (z_1 - 1) \cdot c \\ &\equiv n - c \pmod{N}. \end{aligned}$$

Therefore, $(x, y, z - 1)$ represents the same value $n - c$ as the point $(x_1, y_1, z_1 - 1)$. However, $(x_1, y_1, z_1 - 1)$ was visited before $(x, y, z - 1)$; we see that $(x, y, z - 1)$ is not in the pattern. This contradicts the fact that (x, y, z) is an xy-point. Therefore, (x, y, z) represents the same value as a previously visited point on the xy-plane. Similarly, we can prove that each yz-point (respectively, xz-point) represents the same value as a previously visited point in the pattern on the yz-plane (respectively, xz-plane). It is clear that each x-point is both an xy-point and xz-points. Therefore, each x-point represents the same value as a previously visited point in the pattern on the x-axis. Similarly, each y-point (respectively, z-point) represents the same value as a previously visited point in the pattern on the y-axis (z-axis). Let (x, y, z) be a lattice point not in the pattern but at the corner bounded by three pairwise perpendicular planes, none of which is a coordinate plane. It is clear that this lattice point is an x-point, a y-point, and a z-point simultaneously. Therefore, it represents the same value as a previously visited point in the pattern that is on all three axes. Hence it represents 0. This implies that the resulting pattern has at most one such corner. Figure 3 presents such a "general" pattern.

The proof of the lemma is complete. □

Now we are ready to prove the following theorem, which gives an upper bound for $N(d, 3)$.

THEOREM 3. *Let $d \geq 1$ be any integer. Then we have*

$$N(d, 3) \leq \frac{1}{14 - 3\sqrt{3}} (d + 3)^3.$$

Proof. Let $D = D(\alpha)$ be the solid in the three-dimensional Euclidean space (see Fig. 4) bounded by

$$0 \leq x + y \leq \alpha,$$

$$0 \leq y + z \leq \alpha,$$

$$0 \leq z + x \leq \alpha.$$

The solid D^* consisting of all lattice points in

$$S_0^{(3)} \cup S_1^{(3)} \cup \cdots \cup S_\alpha^{(3)}$$

is contained in D. In fact, D^* is bounded by the planes

$$x = 0, \quad y = 0, \quad z = 0, \quad \text{and} \quad x + y + z \leq \alpha.$$

Let $D + (u, u, u)$ be the solid consisting of all points $(x, y, z) = (x' + u, y' + u, z' + u)$, where (x', y', z') is a point in D. More specifically, $D + (u, u, u)$ is the solid

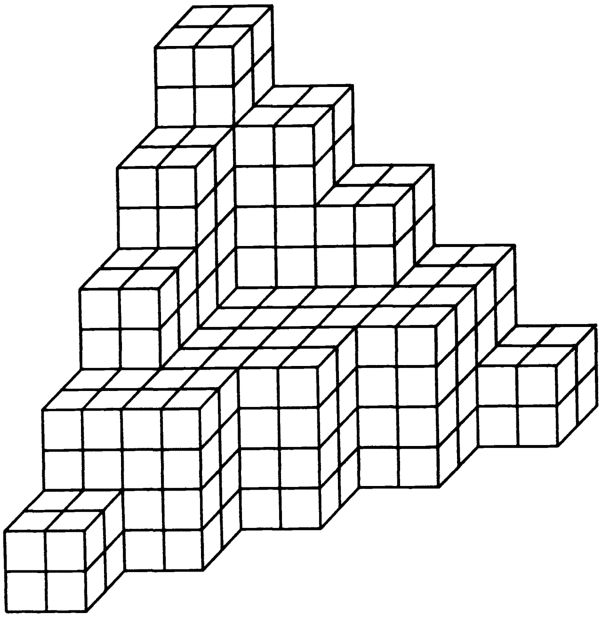


FIG. 3. *The "general" filled pattern in the three-dimensional case.*

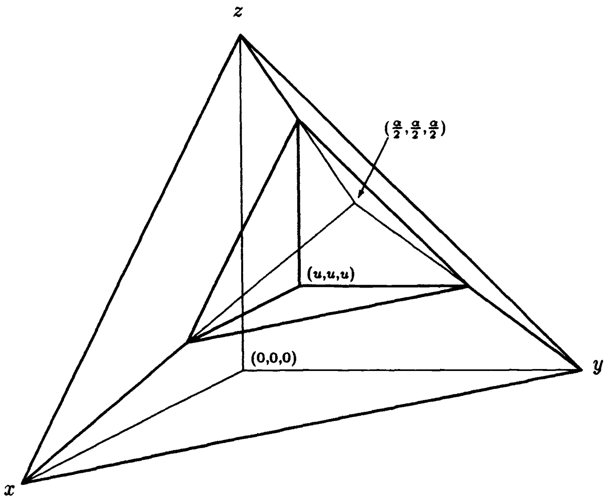


FIG. 4. *The solid* $D_1 = D\backslash(D + (u, u, u))$.

that is isomorphic to D, but the origin is (u, u, u) instead of $(0, 0, 0)$. Let $D_1 = D\backslash (D + (u, u, u))$. It follows from Lemma 2 that, if a lattice point (x_0, y_0, z_0) is blank after the visitation, then all lattice points (x, y, z) with $x \geq x_0$, $y \geq y_0$, and $z \geq z_0$ are also blank. In other words, if (x_0, y_0, z_0) is blank, then all the lattice points in the solid $D + (x_0, y_0, z_0)$ are also blank. Hence the best possible filled pattern resulting from the construction is contained in the solid D_1 for some u. It is clear that $(\frac{\alpha}{2}, \frac{\alpha}{2}, \frac{\alpha}{2})$ is the only vertex of $D = D(\alpha)$ not on any of the coordinate planes. It is also the corresponding vertex of $D \cap (D + (u, u, u))$ since these two solids are similar. We may setup a coordinate system according to $D + (u, u, u)$, which has the origin (u, u, u). Suppose that the vertex $(\frac{\alpha}{2}, \frac{\alpha}{2}, \frac{\alpha}{2})$ has a new coordinate $(\frac{\beta}{2}, \frac{\beta}{2}, \frac{\beta}{2})$ in $D + (u, u, u)$ with respect to its new coordinate system. Then $\frac{\alpha}{2} - u = \frac{\beta}{2}$, i.e., $\beta = \alpha - 2u$. It is easy to see that the volume $V(D)$ of D is

$$V(D) = \frac{\alpha^3}{4}.$$

Let $L(D)$ denote the number of lattice points contained in D. Then

$$\frac{\alpha^3}{4} \leq L(D) \leq \frac{(\alpha + 2)^3}{4}.$$

Therefore, the number of lattice points in D_1 is at most

$$\frac{(\alpha + 2)^3}{4} - \frac{\beta^3}{4} = \frac{(\alpha + 2)^3}{4} - \frac{(\alpha - 2u)^3}{4}.$$

The number of steps to reach other lattice points in D_1 from $(0, 0, 0)$ is at most $3u + \alpha - 2u = u + \alpha$. If d is the diameter of $G(N(d, 3), A)$, then $u + \alpha \leq d$, i.e., $u \leq d - \alpha$. Hence

$$(9) \qquad L(D_1) \leq \frac{(\alpha + 2)^3}{4} - \frac{(3\alpha - 2d)^3}{4}.$$

Denote by $f(\alpha)$ the function of α on the right-hand side of (9). It is easy to see that $f(\alpha)$ reaches its minimum when

$$\alpha = \frac{2 + 2\sqrt{3d}}{3\sqrt{3} - 1}.$$

It therefore follows from (9) that

$$\begin{aligned} N(d, 3) \le L(D_1) &\le f\left(\frac{2 + 2\sqrt{3}d}{3\sqrt{3} - 1}\right) \\ &= \frac{1}{4}\left(\frac{2 + 3\sqrt{3}d}{3\sqrt{3} - 1} + 2\right)^3 - \frac{1}{4}\left(3\cdot\frac{2 + 3\sqrt{3}d}{3\sqrt{3} - 1} - 2d\right)^3 \\ &= \frac{6\sqrt{3}(d+3)^3}{(3\sqrt{3} - 1)^3} - \frac{2(d+3)^3}{(3\sqrt{3} - 1)^3} \\ &= \frac{1}{14 - 3\sqrt{3}}(d+3)^3. \end{aligned}$$

This completes the proof of the theorem. □

3. $N(d, k)$ and the postage stamp problem. In the previous section, we gave an exact formula for $N(d, 2)$. In addition, bounds were given to estimate $N(d, 3)$. In this section, we study $N(d, k)$ for $k \ge 4$. We also establish a striking relation between $N(d, k)$ and a problem in additive number theory.

Let $A_k = \{0 < a_1 = 1 < a_2 < \cdots < a_k\}$ be a set of $k + 1$ integers. A_k is called an *h-basis* for n if every integer in $[0, n]$ can be written as a sum of h elements of A_k, repetitions being allowed, i.e., $hA_k \supseteq [0, n]$. Let $n(h, A_k)$ denote the largest n for which A_k is an h-basis. Let

$$n(h, k) = \max_{A_k} n(h, A_k).$$

The problem of calculating $n(h, k)$ is sometime referred to as *the postage stamp problem*, due to a rather obvious combinatorial interpretation. It has received extensive study in additive number theory. Those interested in this problem may consult Selmer's research monograph [11].

By the definitions of $N(d, k)$ and $n(h, k)$, we have $N(d, k) \ge n(d, k)$. Therefore, any lower bound for $n(d, k)$ would be also a lower bound for $N(d, k)$. The best lower bound for $n(h, k)$ as h tends to infinity is due to Mrose [9], who showed that, for $k \ge 4$,

$$n(h, k) \ge \gamma_k 2^{\lfloor k/4 \rfloor}\left(\frac{h}{k}\right)^k + O(h^{k-1}),$$

where $\gamma_k = 1, 1.024, 1.205$, or 1.388, accordingly as $k \equiv 0, 1, 2$, or $3 \pmod 4$. Therefore, we have the following lower bound for $N(d, k)$.

THEOREM 4. *For $k \ge 4$,*

$$N(d, k) \ge \gamma_k 2^{\lfloor k/4 \rfloor}\left(\frac{d}{k}\right)^k + O(d^{k-1}) \quad \text{as } d \to \infty,$$

where γ_k is defined as above.

In particular, we have

$$N(d,4) \geq \frac{1}{128} d^4 + O(d^3),$$

$$N(d,5) \geq \frac{2.048}{5^5} d^5 + O(d^4).$$

However, the lower bounds in Theorems 1 and 2 cannot be obtained in this way. In fact, Stöhr and several later writers showed that

$$n(h,2) = \left\lfloor \frac{h^2+6h+1}{4} \right\rfloor,$$

and Hofmeister [5], [6] showed that

$$n(h,3) = \tfrac{4}{81}h^3 + O(h^2).$$

On the other hand, we can consider both $N(d, k)$ and $n(h, k)$ as k tends to infinity for any fixed d and h. Mrose [10] proved that

$$n(2,k) \geq \tfrac{2}{7}k^2 + O(k),$$

and Windecker [12] proved that

$$n(3,k) > \tfrac{4}{81}k^3 \quad \text{for all } k.$$

Therefore, from the fact $N(d, k) \geq n(d, k)$, we have the following theorems.

THEOREM 5. *It holds that* $N(2, k) \geq \frac{2}{7}k^2 + O(k)$.

THEOREM 6. *It holds that* $N(3, k) > \frac{4}{81}k^3$.

4. Minimal diameter in distributed loop networks. In §§2 and 3, we calculated the number $N(d, k)$. Given the degree k and diameter d, the problem there is to find a Cayley digraph $G(N, A)$ of $\mathbf{Z}/(N)$ with a maximal order N so that $|A| = k$ and the diameter of $G(N, A)$ is less than or equal to d. Conversely, this problem can be regarded as finding a Cayley digraph $G(N, A)$ so that the diameter $d(G(N, A))$ is equal to $d(N, k) = \min\{d(G(N, A)) \mid |A| = k\}$. Although $d(N, k)$ and $N(d, k)$ were defined for any subset A of $[1, N-1]$, we restrict ourselves to the case in which A contains the element 1 in this paper. However, many techniques and results can be applied to the general case. It is obvious that $d(N, 1) = N - 1$. For $k \geq 2$, bounds were given by Wong and Coppersmith [13] as follows:

$$\sqrt[k]{k!N} - \frac{k+1}{2} \leq d(N,k) \leq k\sqrt[k]{N} - k. \tag{10}$$

They also improved the lower bound in the case where $k = 2$,

$$d(N,2) \geq \lceil \sqrt{3N} \rceil - 2.$$

Let $lb(N, 2) = \lceil \sqrt{3N} \rceil - 2$. The case where $k = 2$ has received extensive study in recent years mainly because Cayley digraphs $G(N, A)$ with $|A| = 2$ are natural generalization of the popular ring network $G(N, s)$. Large infinite families of $G(N, A)$ with $|A| = 2$ have been constructed that have diameters equal to $lb(N, 2)$. Refer to Erdös and Hsu [3], Fiol et al. [4], Hwang and Xu [7], and the recent survey by Bermond, Comellas, and Hsu [1].

For $k = 3$, inequalities in (10) lead to

$$\sqrt[3]{6N} - 2 \leq d(N,3) \leq 3\sqrt[3]{N} - 3.$$

Among other results, Erdös and Hsu [3] improved the upper bound to be of $2.7\sqrt[3]{N}$. In this section, we utilize the results in §§2 and 3 to obtain new bounds for $d(N, k)$.

In the proof of Theorem 2, we actually showed that

$$N(d, 3) \geq \tfrac{1}{16}d^3 + \tfrac{3}{8}d^2 + O(d).$$

Hence $d(N, 3) \leq \sqrt[3]{16N} = 2.5189\sqrt[3]{N}$ for an infinite family of N. By Theorem 3, we have

$$d(N, 3) \geq \sqrt[3]{14 - 3\sqrt{3}}\sqrt[3]{N} - 3 = 2.06486\sqrt[3]{N} - 3.$$

This is true for *every* integer $N \geq 58$. Therefore, we have the following theorem.

THEOREM 7. *Let $G(N, A)$ be the Cayley digraph on $\mathbf{Z}/(N)$ with $1 \in A$. Let*

$$d(N, k) = \min\,\{d(G(N, A)) \,|\, A \textit{ with } |A| = k\},$$

where $d(G)$ is the diameter of the graph G. Then

$$\sqrt[3]{14 - 3\sqrt{3}}\sqrt[3]{N} - 3 \leq d(N, 3) \leq \sqrt[3]{16}\sqrt[3]{N},$$

where the lower bound is for all $N \geq 58$, while the upper bound is true for an infinity family of N.

The lower bound $\sqrt[3]{14 - 3\sqrt{3}}\sqrt[3]{N} - 3$ for $d(N, 3)$ in Theorem 7 is certainly much better than $\sqrt[3]{6N} - 2$. Our lower bound is sharp, as shown in Table 2.

Following Theorem 4, we have the following upper bound for $d(N, k)$ when $k \geq 4$.

THEOREM 8. *For $k \geq 4$ and $d(N, k)$ as defined before,*

$$d(N, k) \leq \frac{k}{\sqrt[k]{\gamma_k \cdot 2^{\lfloor k/4 \rfloor}}} \cdot \sqrt[k]{N}$$

for an infinity family of integers N, where $\gamma_k = 1, 1.024, 1.205,$ or 1.388, accordingly as $k \equiv 0, 1, 2,$ or $3 \pmod 4$.

We note that when $k = 4$, this upper bound becomes $\sqrt[4]{128} \cdot \sqrt[4]{N}$, which is a significant improvement on the result of Wong and Coppersmith in [13].

We conclude this section by providing an example to illustrate our result. In Theorems 2 and 5, let $r = t$ and $d = 4r = 4t$. Then

$$N_1 = \frac{d^3}{16} + \frac{3d^2}{8} + d + 3 = 4t^3 + 6t^2 + 4t + 3,$$

$$b = 3d - 8r + 5 = 4t + 5,$$

$$c = rb + d - 3r + 2 = 4t^2 + 6t + 2.$$

TABLE 2
$lb(N, 3)$ and $d(N, 3)$.

N	$lb(N, 3) = \lfloor\sqrt[3]{14 - 3\sqrt{3}}\sqrt[3]{N}\rfloor - 3$	$d(N, 3)$
78	6	7
111	7	8
138	8	9
176	9	10
217	10	11

TABLE 3
$d(N_1, A)$ *and* $d(N, 3)$.

$d(N_1, A)$	$d(N, 3)$
$d(G(67, \{1, 13, 30\})) = 8$	$d(67, 3) = 7$
$d(G(167, \{1, 17, 56\})) = 12$	$d(167, 3) = 10$
$d(G(371, \{1, 21, 90\})) = 16$	$d(371, 3) = 14$

Let $A = \{1, b, c\}$. Then the network $G(N_1, A)$ has diameter $4t$. In comparison, we list both the exact value $d(N, k)$ and the result from our construction in Table 3.

5. Concluding remarks. In this paper, we study the Cayley digraph $G(N, A)$ associated with $\mathbf{Z}/(N)$ and $A \subset [0, N-1]$. In particular, we initiate a systematic study of $N(d, k)$, which is the maximum N, so that there is an $A = \{a_1, a_2, \ldots, a_k\}$, $1 = a_1 < a_2 < \cdots < a_k$ and that $G(N, A)$ has the diameter $d(G(N, A)) \leq d$. We have obtained exact value for $N(d, 2)$ and upper and lower bounds for $N(d, k)$, $k \geq 3$. Using these results, we have derived new bounds for minimal diameter in the construction of distributed loop networks. An interesting relationship is established between the $N(d, k)$ problem and the postage stamp problem in additive number theory.

In our study, we restrict the subset A of $[0, N-1]$ to be the one with $a_1 = 1$. The digraph $G(N, A)$ constructed having $a_1 = 1$ contains a Hamiltonian circuit. This type of Cayley digraphs (networks) merits further study because $G(N, A)$ is a generalization of the popular ring network that has $A = \{1\}$. Mathematically, A can be any subset of $[0, N-1]$. Moreover, the underlining group of the Cayley digraph can be generalized to an arbitrary finite group. Jia [8] studied some related problems in the general case and proved a relation between Cayley digraphs and bases for finite groups.

It is interesting to further explore the relationship between $N(d, k)$ and $d(N, k)$. Although these two variables are inverse to each other in some sense, they do behave differently in some other manner. For example, $N(d, k)$ is a monotonically increasing function of d when k is fixed. However, $d(N, k)$ is a zig-zag function of N and $|d(N, k) - d(N-1, k)| \leq 1$. Regarding the case where $k = 3$, discussed in this paper, detailed calculation shows that $d(N, 3) = 2, 3, 4, 5, 6$, and 7 when N is in $[5, 9]$, $[10, 16]$, $[17, 27]$, $[28, 40]$, $[41, 57]$, and $[58, 78]$, respectively. When N is in $[79, 111]$, $d(N, 3) = 8$, except for $N \in [105, 110]$ when $d(N, 3) = 9$. The same thing happens for N in $[112, 138]$ and $[139, 176]$, where $d(N, 3) = 9$ and 10, respectively, except for $N = 133$, 137, 172, and 174, where the diameter is one larger than the diameter of network of other orders nearby.

In the case where $k = 2$, the lower bound $\lceil \sqrt{3N} \rceil - 2$ for $d(N, 2)$ obtained by Wong and Coppersmith [13] has been shown to be very closed to $d(N, 2)$ by various authors. The lower bound $lb(N, 3)$ we obtained in this paper for $d(N, 3)$ is a major improvement over $\sqrt[3]{6N} - 2$, as obtained by Wong and Coppersmith [13]. We do not know if there is an infinite family of N for which $lb(N, 3) = d(N, 3)$.

Finally, we note that Theorem 3 can be generalized to the case where $N(d, k)$, $k \geq 4$. This will provide the following lower bound for $N(d, k)$ in the case where $k = 4$:

$$N(d, 4) \geq \tfrac{1}{125} d^4 + O(d^3),$$

which is slightly stronger than the lower bound obtained by using Mrose's result concerning the postage stamp problem.

REFERENCES

[1] J.-C. BERMOND, F. COMELLAS, AND D. F. HSU, *Distributed loop computer networks, A survey*, J. Parallel Distributed Comput., to appear.

[2] D. V. CHUDNOVSKY, G. V. CHUDNOVSKY, AND M. M. DENNEAU, *Regular graphs with small diameter as models for interconnection networks*, in Proc. 3rd Internat. Conference on Supercomputing, Boston, May 1988, pp. 232–239.

[3] P. ERDÖS AND D. F. HSU, *Distributed loop networks with minimum transmission delay*, Theoret. Comput. Sci., 100 (1992), pp. 223–241.

[4] M. A. FIOL, J. L. A. YEBRA, I. ALEGRE, AND M. VALERO, *A discrete optimization problem in local networks and data alignment*, IEEE Trans. Comput., C-36 (1987), pp. 702–713.

[5] G. HOFMEISTER, *Asymptoticsche Abschatzungen für dreielementige Extremalbasen in naturlichen Zahlen*, J. Reine Angew. Math., 232 (1968), pp. 77–101.

[6] ———, *Die dreielementigen Extremalbasen*, J. Reine Angew. Math., 339 (1983), pp. 207–214.

[7] F. K. HWANG AND Y. H. XU, *Double loop networks with minimum delay*, Discrete Math., 66 (1987), pp. 109–118.

[8] X.-D. JIA, *Thin bases for finite nilpotent groups*, J. Number Theory, 41 (1992), pp. 303–313.

[9] A. MROSE, *Ein rekursives Konstruktionverfahren fur Abschnittsbasen*, J. Reine Angew Math., 271 (1974), pp. 214–217.

[10] ———, *Untere Schranken für die Reichweiten von Extremalbasen fester Ordnung*, Abh. Math. Sem. Univ. Hamburg, 48 (1979), pp. 118–124.

[11] E. S. SELMER, *The Local Postage Stamp Problem*, Research Monograph, Department of Mathematics, University of Bergen, Bergen, Norway.

[12] R. WINDECKER, *Eine Abschnittsbasis dritter Ordnung*, Det Kongelige Norske Videnskabers Selskab, 9 (1976), pp. 1–3.

[13] C. K. WONG AND D. COPPERSMITH, *A combinatorial problem related to multimodule memory organization*, J. Assoc. Comput. Mach., 1974, pp. 392–401.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 72–89, February 1994

ON-LINE COLORING AND RECURSIVE GRAPH THEORY*

H. A. KIERSTEAD†‡, S. G. PENRICE†‡, AND W. T. TROTTER†§

Abstract. An on-line vertex coloring algorithm receives the vertices of a graph in some externally determined order, and, whenever a new vertex is presented, the algorithm also learns to which of the previously presented vertices the new vertex is adjacent. As each vertex is received, the algorithm must make an irrevocable choice of a color to assign the new vertex, and it makes this choice without knowledge of future vertices. A class of graphs Γ is said to be on-line χ-bounded if there exists an on-line algorithm A and a function f such that A uses at most $f(\omega(G))$ colors to properly color any graph G in Γ. If H is a graph, let Forb(H) denote the class of graphs that do not induce H. The goal of this paper is to establish that Forb(T) is on-line χ-bounded for every radius-2 tree T. As a corollary, the authors answer a question of Schmerl's; the authors show that every recursive cocomparability graph can be recursively colored with a number of colors that depends only on its clique number.

Key words. on-line algorithm, graph coloring, recursive function

AMS subject classification. 05C15

1. Introduction. The main result of this article can be formulated in terms of recursive function theory or on-line algorithms. Since the on-line formulation gives a slightly stronger statement and is more universally accessible, we adopt it. However, since the roots of the subject lie in recursive graph theory, we begin with a brief summary of results in this area. A *recursive graph* is a countable graph $G = (V, E)$ such that there exist algorithms (Turing machines) for computing the characteristic functions of V and E. A recursive graph is *highly recursive* if each vertex has finite degree and there exists an algorithm for calculating the degree of each vertex. A graph is *recursively k-colorable* if there exists an algorithm that computes a proper k-coloring of the vertices of the graph. The *recursive chromatic number* of a graph is the least k such that the graph is recursively k-colorable. During the 1970s, several authors, including Manaster and Rosenstein [MR] and Bean, Schmerl, and Kierstead studied the recursive chromatic number of various classes of graphs. For example, Bean [B] proved that every planar highly recursive graph has recursive chromatic number at most 6. Schmerl [S1] showed that every highly recursive k-colorable graph can be recursively $2k$-1-colored, and, in general, this is best possible. He also proved [S2] that Brooks's bound on the chromatic number of a graph also holds for the recursive chromatic number of a highly recursive graph. Kierstead [K2] proved that the recursive chromatic number of a highly recursive perfect graph was, at most, 1 more than its chromatic number and that the recursive edge chromatic number of a highly recursive graph was, at most, 1 more than its edge chromatic number.

While there were many positive results for highly recursive graphs, the results for recursive graphs were almost always negative, unless the class of graphs under consideration had bounded degree. For example, Bean [B] showed that there are recursive forests whose recursive chromatic number is infinite. However, Kierstead [K1] did give a positive result for a similar problem. He answered a question of Schmerl, by showing that every recursive ordered set with width w could be partitioned into at most $(5^w - 1)/4$ recursive chains. From a purely graph theoretical point of view, partitioning

* Received by January 13, 1992; accepted for publication (in revised form) February 9, 1993.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

‡ This author was partially supported by Office of Naval Research grant N00014-90-J-1206.

§ This author was partially supported by National Science Foundation grant DMS-87-13994.

an ordered set of width w into θ chains is equivalent to partitioning a comparability graph with independence number w into θ complete subgraphs, which, in turn, is equivalent to properly coloring a cocomparability graph with clique number w using θ colors. However, Kierstead's algorithm made explicit use of the orientation of the recursive ordered set. These considerations led Schmerl to ask whether there exists a function $f(w)$ such that every recursive comparability graph with independence number w can be partitioned into $f(w)$ recursive cliques. Kierstead and Trotter [KT1] showed that this was true for comparability graphs of interval orders.

Now we consider Schmerl's question from the point of view of on-line algorithms. An *on-line graph* is a structure $G^< = (V, E, <)$, where $G = (V, E)$ is a graph, V is finite or countably infinite, and $<$ is a linear ordering of V. (If V is infinite, then $<$ has the order type of the natural numbers.) We call $G^<$ an *on-line presentation* of the graph G. The on-line subgraph of $G^<$ induced by a subset $X \subset V$ is the on-line graph $G^<[X] = (X, E', <')$, where E' is the set of edges in E both of whose endpoints are in X and $<'$ is $<$ restricted to X. Let $V_i = \{x_1, \dots, x_i\}$ denote the first i vertices of V in the linear order $<$ and set $G_i^< = G^<[V_i]$. More generally, we refer to on-line structures $S^<$, where S is some structure such as an ordered set or partitioned graph. An algorithm for coloring the vertices of an on-line graph $G^<$ (or, more generally, calculating some function on the universe of an on-line structure) is said to be *on-line* if the color of a vertex x_i is determined solely by $G_i^<$. Intuitively, the algorithm colors the vertices of $G^<$ one at a time in some externally determined order $x_1, \dots, x_n$, and, at the time a color is irrevocably assigned to the vertex x_i, the algorithm can only see $G_i^<$. A simple but important example of an on-line algorithm is the algorithm First-Fit, denoted by FF, which colors the vertices of G with an initial sequence of the colors $\{1, 2, \dots\}$ by assigning to the vertex x_i the least possible color not already assigned to any vertex $x \in V_{i-1}$ such that x is adjacent to x_i.

Usually, an algorithm for recursively coloring recursive graphs results in an on-line algorithm, while more specialized algorithms for coloring highly recursive graphs do not. The reason for this is that algorithms for coloring highly recursive graphs can learn about the neighbors of a vertex, or the neighbors of the neighbors of a vertex, and so forth, before coloring the vertex. An on-line algorithm for coloring graphs always produces an algorithm for coloring recursive graphs. In this vein, the proof of Kierstead's recursive chain covering theorem actually yields the following slightly stronger statement.

THEOREM 1.1. *There exists an on-line algorithm A that will partition any on-line ordered set $P^<$ into* $(5^w - 1)/4$ *chains.*

Similarly, Bean's example of a forest with infinite recursive chromatic number produces an on-line tree that cannot be finitely colored by any on-line algorithm.

Schmerl's question proves to be a special case of a more general problem. Before continuing, we introduce some terminology and graph theoretical results. The *clique size* and *chromatic number* of a graph G are denoted by $\omega(G)$ and $\chi(G)$, respectively. Let A be an on-line graph coloring algorithm. Then $\chi_A(G^<)$ denotes the number of colors A uses to color the on-line graph $G^<$, and $\chi_A(G)$ denotes the maximum of $\chi_A(G^<)$ over all on-line presentations $G^<$ of G. A class of graphs Γ is said to be *χ-bounded* if there exists a function f such that, for all $G \in \Gamma$, $\chi(G) \le f(\omega(G))$. Easy examples of χ-bounded classes include the class of perfect graphs (which include cocomparability graphs), the class of line graphs, and, more generally, the class of claw-free graphs. Similarly, for an on-line algorithm A, the class Γ is *χ_A-bounded* if there exists a function f such that, for all $G \in \Gamma$, $\chi_A(G) \le f(\omega(G))$. The class Γ is *on-line χ-bounded* if Γ is χ_A-bounded for some on-line algorithm A. The class of perfect graphs is not on-line χ-bounded. In fact, the subclass of trees is not even on-line χ-bounded as we noted above. However, the class of claw-free graphs is on-line χ-bounded. We now rephrase Schmerl's question in these terms.

Question 1.2. Is the class of cocomparability graphs on-line χ-bounded?

A graph $H = (X, F)$ is an *induced subgraph* of a graph $G = (V, E)$ if and only if (1) $X \subset V$, and (2) for all vertices $x, y \in X$, $xy \in F$ if and only if $xy \in E$. For a graph H, let Forb(H) be the class of graphs G such that H is not isomorphic to an induced subgraph of G. In the mid-1970s, Gyárfás [G1] and Sumner [Su] independently formulated the following conjecture.

CONJECTURE 1.3. *For any tree T, the class of graphs* Forb(T) *is χ-bounded.*

Several comments about this conjecture are in order. First, it is easy to show (see [G1] or [Su]) that, if $\chi(G) = k$, then G contains every tree T on k vertices as a subgraph, but not necessarily as an induced subgraph. Second, if Forb(H) is χ-bounded, then H is acyclic. This follows immediately from a result of Erdös and Hajnal [EH] that, for every positive integer $i > 2$, there exists a graph G_i such that both the girth and chromatic number of G_i are at least i. Such graphs have clique number 2 and do not contain any graph that contains a cycle of length i. Finally, if F is a forest, Forb(F) is χ-bounded if and only if, for each of the connected components T_i of F, the class Forb(T_i) is χ-bounded. Thus, if the conjecture is true, its proof yields a characterization of those graphs H such that Forb(H) is χ-bounded.

Rödl proved a weaker version of the conjecture. He showed [KR] that, for every tree T and complete bipartite graph $K_{t,t}$, the class Forb(T) ∩ Forb($K_{t,t}$) is on-line χ-bounded. Gyárfás [G2] showed that the conjecture is true when T is any path. Gyárfás, Szemerédi, and Tuza [GST] verified the conjecture for triangle-free graphs in Forb(T), when T is any radius-2 tree, and Kierstead and Penrice [KP1] extended this result by showing that Forb(T) is χ-bounded whenever T has radius 2. The latter two results use Rödl's theorem. We need the following strengthening of Rödl's result due to Kierstead and Penrice [KP1].

THEOREM 1.4. *For every tree T and complete bipartite graph $K_{t,t}$,* Forb(T) ∩ Forb($K_{t,t}$) *is χ_{FF}-bounded.*

An old result of Chvátal [C] shows that Forb(P_4) is ψ_{FF}-bounded, where P_n is a path on n vertices. Gyárfás and Lehel [GL3] made an exciting and unexpected breakthrough when they extended this result by proving that Forb(P_5) is on-line χ-bounded. They also showed that Forb(P_6) is not on-line χ-bounded. Thus, if Forb(T) is on-line χ-bounded for some tree T, then T has radius at most 2. The central result of this article is that this condition is not only necessary, but is also sufficient.

THEOREM 1.5. *For every tree T, the class* Forb(T) *is on-line χ-bounded if and only if T is a radius-*2 *tree.*

We are indebted to Gyárfás for reminding us that, as a consequence of Gallai's characterization of comparability graphs [Ga], cocomparability graphs do not induce the radius-2 tree obtained by subdividing each edge of $K_{1,3}$ (see Fig. 1.1). Of course, this is part of the easy direction of Gallai's characterization and can be readily verified from scratch. Thus, as an immediate corollary, we obtain the following answer to Schmerl's question.

COROLLARY 1.6. *The class of cocomparability graphs is on-line χ-bounded.*

This paper is organized as follows. In the remainder of this section, we state some preliminary results and review our notation and terminology. In §2 we give an overview of the off-line proof and the problems we must deal with to create an on-line algorithm. In §§4 and 5 we develop some purely combinatorial lemmas needed to verify the correctness of the main algorithm. In §5 we also present a key on-line subroutine and in §6

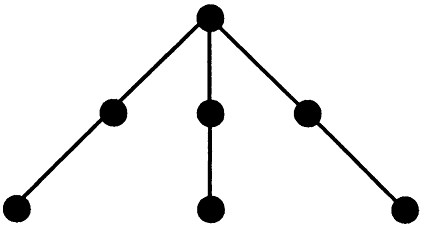


FIG. 1.1.

we give the proof of Theorem 1.5, after giving a more technical reformulation of it as Theorem 6.1.

We use the following easy lemma, which follows, for example, from Turan's theorem.

LEMMA 1.7. *Let D be a directed graph, where Δ^{out} denotes the maximum outdegree of D and where ν and ε denote the number of vertices and edges of D, respectively. Then D contains an independent set of size at least $\nu/(2\Delta^{\text{out}}+1)$.*

Let $R[a, b]$ denote the Ramsey function with the property that every graph on $R[a, b]$ vertices contains an independent set of size a or a complete subgraph of size b.

Let $T_{a,b}$ be the radius-2 tree whose root r is adjacent to a level-1 vertices $x_1, \ldots, x_a$, and each level-1 vertex x_i is adjacent to b leaves $y_{i,1}, \ldots, y_{i,b}$. We call the set $\{y_{i,1}, \ldots, y_{i,b}\}$ a level-2 group. We abbreviate $T_{k,k}$ by T_k. The complete s-partite graph with w vertices in each part is denoted by K^s_w.

If two vertices x and y are adjacent, we write $x \sim y$. If X and Y are sets of vertices such that every vertex in X is adjacent to every vertex in Y, then we write $X \sim Y$. We denote the neighborhood of a vertex x by $N(x) = \{y: x \sim y\}$.

2. Overview. In this section, we give an overview of the proof of the off-line version of our main theorem and the additional problems that must be solved to prove the on-line theorem. This section also serves as a guide to reading the rest of the paper. We begin by noting that, if T' is a subtree of the tree T, then Forb(T') $\subset$ Forb(T), and thus Forb(T') is on-line χ-bounded if Forb(T) is. Thus it suffices to prove that, for all k, Forb(T_k) is on-line χ-bounded. In §3, two elementary combinatorial lemmas on trees are presented, which simplify the remaining arguments. In particular, quasi-induced trees are defined, and it is shown that it suffices to prove that the smaller class q Forb(T_k) of graphs that do not contain a quasi-induced T_k is on-line χ-bounded.

The central idea in both the on-line and off-line proofs is the notion of s-templates and their use to partition the graph so that the vertices can be properly colored in terms of local and global colors. Roughly, an s-template is a complete s-partite graph with a very large number of vertices in each part. The exact definition of its part size depends on $t = \omega(G)$ and s. However, there will be an absolute upper bound on part size in terms of t. At a given stage of a double induction on t and s, we assume the following for some bound c depending on s and t:

(1) If H is a graph on q Forb(T_k) such that either (a) $\omega(H) < t$ or (b) $\omega(H) = t$ and H does not contain an induced s-template, then $\chi(H) \le c$.

It is easy to partition the vertices of G as $(X, B_1, O_1, B_2, O_2, \ldots, B_n, O_n)$ so that, for $1 \le i < j \le n$, the following assumption holds:

(2) (a) B_i is an s-template,

(b) Each vertex in O_i is adjacent to at least k vertices in some part of B_i,

(c) No vertex in B_j is adjacent to k or more vertices in any part of B_i,

(d) X does not contain an s-template.

By part (b) of assumption (1), we can color X with c colors. Using a different disjoint set of $c|B_i|$ colors, we can, by part (a), color each O_i with $c|B_i|$ colors so that all vertices of O_i that receive the same color have a common neighbor in B_i. However, two adjacent vertices, one in O_i and the other in O_j with $i \neq j$, may receive the same color. Finally, using a third disjoint set of $|B_i|$ colors, we can color each B_i so that each vertex of B_i receives a distinct color. Again, two adjacent vertices, one in B_i and the other in B_j with $i \neq j$, may receive the same color. Call these colors local colors and let $\langle \alpha \rangle$ denote the set of vertices with local color α.

It remains to show that we can color each local color class $\langle \alpha \rangle$ with a bounded number of colors. Before describing this coloring, we must mention one technical complication. A point x is said to be an extra point for an s-template B_i if, roughly, x is adjacent to almost every vertex in every part of B_i. Extra points will create all sorts of minor problems, which require special attention. Fortunately, if any template has too many extra points, we will be able to start over using an algorithm based on $(s + 1)$-templates. For the rest of this informal discussion, we ignore the possibility of extra points. With this sluff, we can give a simple statement of the following crucial properties of our partition. In §4 these properties are stated in full technical detail and proved.

(3) There exists a constant d such that, for any vertex x, local color α, and integer $i < n$,

(a) $|\{j : x \sim y, \text{ for some } y \in B_j\}| \leq d$,

(b) $|\{j : x \text{ is adjacent to at least } k \text{ vertices in } O_j \cap \langle \alpha \rangle\}| \leq d$,

(c) $|\{j : \text{for some } x \in O_i \cap \langle \alpha \rangle, x \text{ is adjacent to at least } k \text{ vertices in } O_j \cap \langle \alpha \rangle\}| \leq d$.

Properties (a)–(c) in the above assumption allow us to color each $\langle \alpha \rangle$ with a bounded number of colors as follows. If α is a local color that is used on a vertex in some B_i, then α is used on exactly one vertex of each B_j. Thus, by part (a), the degree of $\langle \alpha \rangle$ is bounded by d, and so $\langle \alpha \rangle$ can be $d + 1$ colored. Suppose that α is a local color that is used on a vertex of O_i. We define a directed auxiliary graph G' on the vertices $\{1, \ldots, n\}$ by $i \to j$ if and only if there exists a vertex $x \in O_i \cap \langle \alpha \rangle$ such that x is adjacent to at least k vertices of $O_j \cap \langle \alpha \rangle$. By part (c), G' has outdegree at most d and thus can be colored with at most $2d + 1$ colors. We assign each vertex $x \in \langle \alpha \rangle$ a two-coordinate global color. The first coordinate is the color of i in G', if $x \in O_i$. Now let $\langle \alpha, \beta \rangle$ be the subset of $\langle \alpha \rangle$ of vertices whose global color has first coordinate β. By (a) and (b), the degree of $\langle \alpha, \beta \rangle$ is less than d^2, and thus $\langle \alpha, \beta \rangle$ can be colored with d^2 colors. We have thus properly colored G with a bounded number of colors.

Next, we consider the problems involved in implementing the above proof on-line. The major problem is that we cannot calculate the partition of G into $(X, B_1, O_1, B_2, O_2, \ldots, B_n, O_n)$ on-line. When a vertex x is presented, it may appear to belong to the first part X of the partition, but later we may learn that it must be assigned to some B_i or O_i. Without being able to properly assign x to a part of the partition, we have no basis for coloring x. Now suppose that we are very fortunate and that whenever a new vertex x is presented we correctly guess its proper position in the partition. There is still a minor problem. Suppose that x is correctly assigned to O_i. Then x is given a global color, which is based in part on the auxiliary graph G'. However, the presentation of x may cause an edge from i to j to appear in G', when previously we had assigned i and j the same color. This problem is solved in §5, where the definition of the auxiliary graph is slightly modified to facilitate dealing with the extra points. In particular, the auxiliary graph will not be directed.

To handle the more serious problem, our on-line algorithm will maintain a partition of G into $P = (X, B_1, O_1, B_2, O_2, \ldots)$, which approximates the desired partition in the following sense. At any stage $i+1$, when we consider the vertex x_{i+1}, P will be a partition of G_i, which satisfies parts (a)–(d) of assumption (2). At any stage, vertices may be removed from X to form B_{j+1}, where B_j is the last template in the previous partition. When this happens, other vertices may be moved from X to O_i. Once a vertex is assigned to a part of the partition other than X, it will never move. Thus, when we are presented with x_{i+1}, we try to assign x_{i+1} to some O_j. If this is not possible, we try to form a new s-template with x_{i+1}, together with some vertices from X. If this is not possible, we assign x_{i+1} to X.

We are left with the problem of coloring the newly presented vertex x_{i+1}. There is no problem if x_{i+1} is assigned to a new s-template B_j, and, if x is assigned to some O_j, it is relatively easy to color x using the techniques of §5, referred to above. The main problem arises when x_{i+1} is first assigned to X. In this case, we do not know where x_{i+1} will end up, so we must somehow hedge our bets. We would like to color x_{i+1} using the on-line version of part (b) of (1). The set of points currently in X does not contain an s-template, but this property is maintained artificially by removing vertices that would otherwise form an s-template in X. Simply removing vertices from X does not solely solve the problem of coloring vertices of X', the set of points originally assigned to X, because the color of a vertex originally assigned to X continues to influence future vertices even after the point is removed from X. We cannot afford to change the set of colors used for vertices entering X every time vertices are removed from X, because we may have to change color sets an unbounded number of times due to the fact that the template sequence may be unbounded in length. Thus, we want part (b) of (1) to forbid, not s-templates in X, but much larger s-partite graphs K_w^s in X', and we will be able to do so because of the other properties of the algorithm. By parts (a) and (b) of assumption (3), the vertices of a supposed K_w^s in X' could not have been used to form too many different s-templates, nor could this set of vertices intersect too many of the O_j's. Thus, if X' contains a copy B' of K_w^s, then, by setting w large enough, we can assume that, for each part Q_a of B', there exists a large subset Q'_a such that either $Q'_a \subset X$ or $Q'_a \subset O_{j(a)} \cup B_{j(a)}$ for some $j(a)$. Here, large means the size of a part in an s-template. This motivates the intricate induction hypothesis presented in §6, where we will color X' on-line with a two-coordinate color. The first coordinate will ensure that, if $x, y \in X'$, $x \sim y$, and $x \in X' - X$ at the time y is presented, then x and y receive different colors. From this fact, following the remarks above, we will show that no first coordinate color class of X' contains K_w^s for appropriately chosen w. Thus we will be able to color each of these first coordinate color classes on-line by a revised version of the induction hypothesis (b) of (1).

3. Lemmas on radius-2 trees. In this section, we develop some preliminary results about trees. We begin with some fundamental definitions. A graph H is called a *pseudo-induced* $T_{a,b}$ if H has a spanning tree T that is isomorphic to $T_{a,b}$ and the root of T is not adjacent in H to any leaf of T. A graph H is called a *quasi-induced* T_k if H has a spanning tree T that is isomorphic to T_k, and, if xy is an edge in H that is not present in T_k, then x and y are either both level-1 vertices, or they are both level-2 vertices. Note that quasi-induced T_k is the stronger of the two: Every quasi-induced T_k is a pseudo-induced T_k, but a pseudo-induced T_k may have "extra" edges between the first and second levels. In a similar vein, we say that H is an *augmented* K_w^s if $V(H)$ can be partitioned into s sets $Q_1, \ldots, Q_s$ of size w such that $x \sim y$ whenever $x \in Q_i$, $y \in Q_j$, and $1 \le i < j \le s$. We abuse standard usage by calling the Q_i's "parts" of H, even though they need not be independent sets.

We introduced the definition of a quasi-induced T_k to simplify future arguments. It is easier to verify that a graph contains a quasi-induced T_k than it is to verify that a graph contains an induced T_k, because, in the latter case, we must in effect check every pair of vertices to see if the proper edge or nonedge is present, whereas, with quasi-induced trees, we need only check pairs of vertices from different levels. The next lemma shows that the results we desire for graphs that do not contain induced trees can be obtained from results about graphs that do not contain quasi-induced trees.

LEMMA 3.1. *For all positive integers k and t, there exists a positive integer $k'(k, t)$ such that, if $k' = k'(k, t)$ and H is a quasi-induced $T_{k'}$, then H contains either an induced T_k or a clique on t vertices.*

Proof. If k' is sufficiently large, then, by repeated applications of the Bipartite Ramsey Theorem, we may assume that, between any two level-2 groups, either all edges are present, or no edges are present. If there are $R[R[k, t], t]$ level-2 groups, then either there exist t level-2 groups with all edges between any two distinct groups present, in which case there exists a t-clique in H, or there exist $R[k, t]$ level-2 groups with no edges present between any two groups. By Ramsey's theorem, among the level-1 vertices associated with each of these $R[k, t]$ level-2 groups, there exists either a t-clique or a set of k independent vertices. In the latter case, however, we have an induced T_k. □

We use q Forb(T_k) to denote the class of graphs that do not contain a quasi-induced T_k. Lemma 3.1 may then be paraphrased as "If k' is sufficiently large, then q Forb($T_{k'}$) $\supseteq$ Forb(T_k)." Henceforth, our arguments will refer to quasi-induced trees rather than induced trees. Also, we denote by q Forb(K_w^s) the class of graphs that do not contain an augmented K_w^s, and q Forb(T_k, K_w^s) denotes the class of graphs that contain neither a quasi-induced T_k nor an augmented K_w^s.

LEMMA 3.2. *Let $H \in q$ Forb(T_k) be a graph that is spanned by T, a pseudo-induced $T_{k(2a+1),kb}$. Then some level-1 vertex x of T has b neighbors in a distinct level-2 group, other than the level-2 group associated with x.*

Proof. Let $x_1, \ldots, x_{k(2a+1)}$ be the level-1 vertices of T and let $S_1, \ldots, S_{k(2a+1)}$ be the level-2 groups of T. Suppose that no level-1 vertex has b neighbors in a distinct level-2 groups. Define a directed graph D on the vertex set $\{1, \ldots, k(2a + 1)\}$ by directing an arc form i to j, $i \neq j$, if and only if x_i has b neighbors in S_j. By our supposition, $\Delta^{\text{out}}(D) \leq a$. It follows from Lemma 3.2 that D has an independent set of size k. This means that we may assume without loss of generality that, for $i = 1, \ldots, k$, x_i has less than a neighbors in each S_j, $1 \leq j \leq k, j \neq i$. Thus, by removing at most $(k - 1)(a - 1)$ vertices from each S_j, leaving at least k vertices in each level-2 group, we find a quasi-induced T_k, a contradiction. □

4. Definitions and structural lemmas.

Let k, p, s, and t be fixed positive integers such that $s \geq 2$ and $p > k^4$. If B is a subgraph of G isomorphic to an augmented K_w^s with $w = p + k^4$, then we call B a *template* of G. We call a vertex $x \notin B$ an *extra point* of B if x has less than k^4 nonneighbors in each part of B. We call a vertex $x \notin B$ a *strong 1-neighbor* of B if x has k neighbors in some part of B. Note that an extra point is also a strong 1-neighbor. A sequence of templates $B_1, \ldots, B_r$ is called an *acceptable template sequence* if, whenever $1 \leq m < n \leq r$ and $x \in B_n$, x is not a strong 1-neighbor of B_m.

This section has two goals: to define an i-tube and to prove a technical lemma, Lemma 4.1, which will play a crucial role when we verify that our on-line algorithm uses a bounded number of colors. In particular, it will help us to verify that the auxiliary graphs have bounded degree, and it will have other applications as well. The following definition and the statement of Lemma 4.1 are, in fact, the only elements of this section used in the rest of the paper. The first-time reader may wish to proceed to §5 after studying the remarks following the statement of Lemma 4.1 and return to the proofs of this section later.

DEFINITION. Let i be a nonnegative integer. Then we call a subgraph $U = (B, N_1, N_2, \ldots, N_i)$ an *i-tube* if

(i) B is a template,
(ii) $B \cap N_m = \emptyset$ for $1 \leq m \leq i$,
(iii) $N_m \cap N_n = \emptyset$ for $1 \leq m < n \leq i$,
(iv) $|N_m| = k$ for $1 \leq m < i$ and $|N_i| = 1$,
(v) If $x \in N_1$, then x has at least k neighbors in some part of B,
(vi) $N_m \sim N_{m+1}$ for $1 \leq m \leq i - 1$.

We refer to the unique element of N_i as the *top* of U. We refer to N_m as the mth *level* of U for $1 \leq m \leq i$. Also, B is the 0th level of U. If U is a 0-tube, then B is the top level.

LEMMA 4.1. *There exists a function $f_1(i, k, \rho)$ such that, for every graph $G \in q\,\mathrm{Forb}(T_k)$ and every vertex x of G, if*

(i) $\mathbf{U} = \{U_1, \ldots, U_j\}$ *is a collection of pairwise disjoint i-tubes in G with* 0*-levels $B_1, \ldots, B_j$, forming an acceptable template sequence,*
(ii) *x is not an extra point of any template in the sequence $B_1, \ldots, B_j$,*
(iii) *x has a neighbor in the top level of each of $U_1, \ldots, U_j$,*
(iv) *B_n has less than ρ extra points for $1 \leq n \leq j$,*

then $j < f_1(i, k, \rho)$.

Lemma 4.1 has simple interpretations in the cases where $i = 0$ and $i = 1$. Namely, if x is a vertex of a graph $G \in q\,\mathrm{Forb}(T_k)$ and there exists an acceptable sequence of templates $B_1, \ldots, B_j$ such that x has a neighbor in B_n but x is not an extra point of B_n for $1 \leq n \leq j$, then $j < f_1(0, k, \rho)$, because a template is, in fact, a 0-tube. Similarly, if, instead of assuming that x has neighbors in the B_n's, we assume that x has neighbors $x_1, \ldots, x_j$ such that x_n is a strong 1-neighbor of B_n for $1 \leq n \leq j$, then $j < f_1(1, k, \rho)$, because $B_n \cup \{x_n\}$ is a 1-tube for $1 \leq n \leq j$. (In the latter case, we retain the assumption that x is not an extra point of the templates; for the x_n's, we need not make a distinction between extra points and nonextra points.) The case where $i = 2$ is also used in our proof, but describing it informally outside the context of the algorithm is awkward.

We surmise that tubes play a role in proving off-line results concerning trees of radius larger than 2; this is the theoretical reason for proving a general version of Lemma 4.1 (for arbitrary i) when only the cases where $i = 0$, 1, and 2 are used. (As a practical matter, the heart of the proof, Lemma 4.3, is proved by induction, so the general result is obtained with more economy than proving these three cases separately.) We also remark that the hypotheses of Lemma 4.1 can be weakened if a corresponding change is made in the definition of an i-tube. Specifically, the tubes need not be completely disjoint, provided that the bases form an acceptable template sequence, which, by definition, consists of disjoint templates. However, to realize this apparent strengthening of the lemma, we must add to the definition of an i-tube the condition that no vertex in the tube is an extra point of the base. After making this adjustment, a different version of Lemma 4.1 enables us to prove Theorem 6.1 using an on-line algorithm, which, while more complicated, appears to use fewer colors than the algorithm of this paper.

Establishing Lemma 4.1 requires some other purely technical lemmas. We state and prove each separately.

LEMMA 4.2. *Let B be a template and suppose that $x \notin B$ is not an extra point of B. If $1 \leq q \leq k^4$ and x has q neighbors in some part of B, then there exist vertices $x_1, \ldots, x_q$ in one part P of B and $y_1, \ldots, y_{k^4}$ in a different part Q such that x is adjacent to x_m for $1 \leq m \leq q$ and x is not adjacent to y_m for $1 \leq m \leq k^4$.*

Proof. Since $p \geq k^4$, every part of B contains either k^4 neighbors of x or k^4 nonneighbors of x, by the pigeonhole principle. Since x is not an extra point of B, then some part, say Q', of B contains k^4 nonneighbors of x. By the hypothesis, some part of B, say P', contains q neighbors of x. If $P' \neq Q'$, set $P = P'$ and $Q = Q'$, and we are done. If $P' = Q'$, then, since $s \geq 2$, we may consider another part, say R. As we observed earlier, either R contains k^4 neighbors of x or R contains k^4 nonneighbors of x. In the former case, set $P = R$ and $Q = Q'$. In the latter case, set $P = P'$ and $Q = R$. It is then easy to find the desired vertices. □

As we indicated at the beginning of this section, one of our short-term goals is to prove Lemma 4.1, which states roughly that there is a bound on the number of tubes in which a point can have neighbors. To prove this bound, it is helpful to prove a bound for a sequence of tubes with special properties and then extend the results to more general sequences of tubes. Thus we are motivated to introduce the following definition. A sequence of i-tubes $U_1, \ldots, U_j$, where $U_n = (B_n, N_{1,n}, \ldots, N_{i,n})$, is called an *acceptable j-sequence of i-tubes* if

(i) $U_m \cap U_n = \emptyset$ for $1 \leq m < n \leq j$,

(ii) The sequence $B_1, \ldots, B_j$ is an acceptable template sequence,

(iii) If $x \in U_m$, then x is not an extra point of B_n for $1 \leq m \neq n \leq j$.

Having introduced this definition, we now show that it is reasonably easy to establish the kinds of bounds we seek when we consider acceptable sequences of i-tubes. The following argument, with weaker bounds, was useful in [GST] and [KP1].

LEMMA 4.3. *Let $\{b_i\}$ be the sequence of functions defined by $b_0(k) = k(2k+1)$ and $b_{i+1}(k) = k(2b_i(k)+1) + b_i(k)$ if $i > 0$. Let $\mathrm{U} = \{U_1, \ldots, U_j\}$ be an acceptable j-collection of i-tubes, for some fixed $i \geq 0$, in some graph $G \in q\,$*Forb*(T_k). Suppose that x is a vertex such that x is adjacent to the top vertex of U_n and x is not an extra point of B_n, for $1 \leq n \leq j$. Then $j < b_i(k)$.*

Proof. Let k be fixed and assume for notational ease that $b_i = b_i(k)$ for $i \geq 0$. We induct on i. Let $i = 0$ and suppose for contradiction that x has a neighbor in B_n for $1 \leq n \leq j = b_0$. Our first goal is to find a vertex y, a strictly increasing function σ, and sets $B'_{\sigma(i)}$ and $B''_{\sigma(i)}$ for $i = 1, \ldots, k$ such that

(1) $B'_{\sigma(i)}, B''_{\sigma(i)} \subset B_{\sigma(i)}$,

(2) $|B'_{\sigma(i)}|, |B''_{\sigma(i)}| \geq k^3$,

(3) y is adjacent to every vertex of $B'_{\sigma(i)}$,

(4) y is nonadjacent to every vertex of $B''_{\sigma(i)}$,

(5) $B'_{\sigma(i)} \sim B''_{\sigma(i)}$.

By applying Lemma 4.2 to x and to each of the templates $B_1, \ldots, B_j$, we may find T_1, a pseudo-induced $T_{a,b}$ with $a = b_0$, $b = k^4$, whose level-2 groups are contained in distinct templates. Applying Lemma 3.2, we find a level-1 vertex y of T_1, which has a set $B'_{\sigma(i)}$ of k^3 neighbors in each of k templates $B_{\sigma(1)}, \ldots, B_{\sigma(k)}$, none of which are at the base of the tube containing y. Observing that (by the definition of an acceptable sequence of i-tubes) y is not an extra point of any of the templates, we may apply Lemma 4.2 again to find $B''_{\sigma(1)}, \ldots, B''_{\sigma(k)}$ as desired. Without loss of generality, $\sigma(i) = i$, for all i.

Having found the structures with properties (1)–(5), we now seek to find a quasi-induced T_k. To do so, we construct a sequence of sets $S_1, \ldots, S_k$ and a sequence of vertices $R_1, \ldots, R_k$ as follows. Let R_k be any vertex of B'_k and let S_k be any k-element subset of B''_k. Suppose that $R_k, S_k, R_{k-1}, S_{k-1}, \ldots, R_r, S_r$ are constructed for $r > 1$. Define R_{r-1} to be any vertex of B'_{r-1} that is not adjacent to any vertex of $S_r \cup \cdots \cup S_k$. This is possible, since each vertex w of $S_r \cup \cdots \cup S_k$ has fewer than k neighbors in B'_{r-1} and $|S_r \cup \cdots \cup S_k| < k^2$. Define S_{r-1} to be any k-element subset of B''_{r-1} that does not contain any neighbors of $R_r, \ldots, R_k$. Again, this is possible because each vertex $R_r, \ldots, R_k$ has fewer than k neighbors in B''_{r-1} and $|\{R_r, \ldots, R_k\}| < k$.

When we have chosen $R_1, \ldots, R_k, S_1, \ldots, S_k$, these vertices, together with y, form a quasi-induced T_k.

Suppose that $i > 0$. Without loss of generality, x has a neighbor in each of $N_{i,1}, \ldots, N_{i,b_i}$; say these neighbors are $x_1, \ldots, x_{b_i}$. By induction, we may assume that x has no neighbors in $N_{i-1,1}, \ldots, N_{i-1,b_i-b_{i-1}}$. Then x is the root of a pseudo-induced $T_{a,b}$ with $a = k(2b_{i-1} + 1)$ and $b = k$. Then, by Lemma 3.2, we find a level-1 vertex x_n that has a neighbor in b_{i-1} level-2 groups contained in tubes that do not contain x_n. This is a contradiction of the induction hypothesis, since, by the definition of an acceptable sequence, x_n is not an extra point of any template that is at the base of a tube (from the acceptable sequence) not containing x_n. □

Proof of Lemma 3.1. Let $f_1(i, k, \rho) = (b_i(k))(2\rho + 1)$. Suppose that conditions (i)–(iv) of the hypothesis hold with $j = f_1(i, k, \rho)$. Define a digraph on $\{1, \ldots, j\}$ by directing an arc from m to n if and only if U_n contains an extra point of B_m. Since the U's are pairwise disjoint, the digraph has outdegree at most ρ; by Lemma 1.7, we must find an independent set of size $b_i(k)$ in the digraph. However, independent sets in this digraph are acceptable tube sequences in G. Thus, we may assume that $\mathbf{U}' = \{U_1, \ldots, U_b\}$ is an acceptable b-sequence of i-tubes, where $b = b_i(k)$. Then, since x has a neighbor in the top of each of these tubes, Lemma 4.3 implies that $b_i(k) < b_i(k)$, a clear contradiction. □

5. Lemma for using auxiliary graphs on-line. We now deal with the minor problem mentioned in our overview. We wish to show that to color a graph $G \in q\ \mathrm{Forb}(T_k)$ on-line with a bounded number of colors, it suffices to partition G on-line into independent sets $I_1, \ldots, I_r$ with the following properties:

(a) For all m, $1 \leq m \leq r$, $|\{n\colon I_m \cup I_n \text{ contains an augmented } K_{e,e}\}| \leq f$,

(b) If $H \cong K_{d,d}$ is a subgraph of G, then there exist I_p and I_q such that some subset of $(i_p \cup I_q) \cap H$ contains an augmented $K_{e,e}$.

The number of colors used will be bounded in terms of d, e, f, and $\omega(G)$. Many of the essential ideas are present in a transparent way when we consider the off-line setting. We begin with this argument.

PROPOSITION 5.1. *There exists a constant c depending only on d, e, f, and $\omega(G)$, such that, if $G \in q$* Forb(T_k) *can be partitioned into independent sets as in* (a) *and* (b), *then* $\chi(G) \leq c$.

Proof. Suppose that G is partitioned into independent sets $I_1, \ldots, I_r$ in a way satisfying (a) and (b). Define a graph G' on the sets $I_1, \ldots, I_r$ by declaring that I_m is adjacent to I_n if and only if $\mathrm{I}_m \cup I_n$ contains an augmented $K_{e,e}$. Note that, by the definition of (a), $\Delta(G') \leq f$, and it easily follows that $\chi(G') \leq f + 1$. Let g' be an optimal coloring of G'. Define a two-coordinate coloring g of G as follows. Compute the first coordinate of each vertex x by assigning x the color $g'(I_n)$, where I_n is the independent set in the partition that contains x. After all the first coordinates have been computed, for every $\alpha \in \mathrm{range}\ (g')$, define a graph G_α as the subgraph of G induced by vertices that received color α in their first coordinate. Let g_α be an optimal coloring of G_α and let $g_\alpha(x)$ be the second coordinate of $g(x)$. It is clear that g is a proper coloring of G and that

$$|\mathrm{range}\ (g)| \leq |\mathrm{range}\ (g')|\ \max\ \{|\mathrm{range}\ (g_\alpha)|\colon \alpha \in \mathrm{range}\ (g)\}.$$

It remains for us to verify that $|\mathrm{range}\ (g)|$ is bounded in terms of d, e, f, and $\omega(G)$. We have already noted that $|\mathrm{range}\ (g')| \leq f + 1$. Since $q\ \mathrm{Forb}(T_k, K_{d,d})$ is χ-bounded (indeed, it is χ_{FF}-bounded), it suffices to show that $G_\alpha \in q$ Forb($K_{d,d}$). Suppose that G_α contains an augmented $K_{d,d}$; call this subgraph H. By condition (b), there exist independent sets I_p and I_q of the partition such that $(I_p \cup I_q) \cap H$ contains an augmented $K_{e,e}$. Then, however, I_p and I_q are adjacent in G', so vertices on opposite sides of the $K_{d,d}$ received different colors in their first coordinate of g, contradicting the fact that every vertex of H received color α in its first coordinate. □

We refer to G' as the *auxiliary graph*. To avoid confusion with the vertices of G, we call the points of G' *nodes*. There are several adjustments that must be made in the on-line case. The greatest difficulty arises from the fact that, when G and the partition $I_1, \dots, I_r\}$ are presented on-line in G', it is possible that an edge may be "discovered" in G' well after both of its nodes have appeared. This is because the nodes of the auxiliary graph are sets of vertices of G, and edges are formed in the auxiliary graph only when a large number of edges (of G) are present between the two sets of the partition. Thus the auxiliary graph does not strictly fall under the on-line model. However, as colorers, we are aided by the fact that (again, in contrast to the standard on-line model) we may change the color of a node as we discover new edges incident on the vertex, provided that the number of times we change the color of a node is no larger than the degree of the node.

LEMMA 5.2. *There exists an on-line algorithm A and a constant c depending on d, e, f, and ω such that, if $G^<$is an on-line presentation of a partitioned graph $G = V, E, I_1, I_2, \dots)$, with $\omega(G) \leq \omega$ and $G \in q$* Forb(T_k), *satisfying* (a) *and* (b), *then A colors $G^<$ using at most c colors.*

Proof. Let G' be defined as in Proposition 5.1. Without loss of generality, we may assume the node set of G' is the set of positive integers. As a new vertex in G is presented, it may cause a previously unseen edge to appear in G'. Despite this complication, we seek to maintain a proper coloring of G', even though we may occasionally change the color of a node of G'.

Suppose that a vertex x enters G and is assigned to I_m. If x does not produce any new edges incident on I_m, do not change the coloring of G'. If x does produce one or more new edges incident on I_m, assign I_m a new color (if necessary) so that the color of I_m is different from all of its previous colors, as well as the current colors of all the neighbors of I_m in G'. Since I_m has at most f neighbors in G' and each vertex of G' has at most $f + 1$ different colors throughout its history (since only the addition of an incident edge can cause a color change), a color will always be available for I_m, provided that we use $f^2 + f + 1$ colors to color G'. Moreover, we have something more than a proper coloring: After an edge $I_m I_n$ appears, I_m is never given a color previously held by I_n, and vice versa.

Now x will be assigned a two coordinate color. The first coordinate will be the color of I_m in G' after any edges in G' caused by adding x to G have been added to G'. To compute the second coordinate, apply First-Fit to the subgraph of G induced by the vertices that received the same first-coordinate color as x.

Clearly, this algorithm gives a proper coloring of G, and we have already determined that at most $f^2 + f + 1$ colors are needed in the first coordinate. Thus it suffices to show that the number of colors used in the second coordinate is bounded by a function of d, e, T, and $\omega(G)$. In fact, we have already done so when we argued for Proposition 5.1, because the bound in the second coordinate of that coloring could be realized by applying First-Fit. □

6. The main theorem. In this section, we prove the following technical reformulation of Theorem 1.5, our central result.

THEOREM 6.1. *For every positive integer k, there exists an on-line coloring algorithm A_k and a function c_k such that, if $G^<$ is an on-line presentation of $G \in$ Forb(T_k), then A_k gives $G^<$ a proper coloring using at most $c_k(\omega(G))$ colors.*

Proof. Let k be a fixed positive integer. We have already observed that it suffices to prove the theorem for q Forb(T_k). Another simplifying observation is that it suffices to prove the following, apparently weaker, statement:

(*) For every positive integer t, there exists an on-line algorithm $A_{k,t}$ and an absolute constant $c_{k,t}$ such that, if $G \in q$ Forb(T_k) and $\omega(G) \le t$, then $A_{k,t}$ colors G using at most $c_{k,t}$ colors.

Statement (*) is, in fact, no weaker than Theorem 6.1; we simply use pairwise disjoint sets of colors for each algorithm $A_{k,t}$. If $G^<$ is an on-line presentation of a graph $G \in q$ Forb(T_k), then, whenever $\omega(G_i^<) = \omega(G_{i-1}^<) = t$, we may color x_i using $A_{k,t}$. On the other hand, if $\omega(G_i^<) = \omega(G_{i-1}^<) + 1$, we may begin using $A_{k,t+1}$ and a new set of colors. We then have Theorem 6.1 by taking $c_k(\omega(G)) = \sum_{t=1}^{\omega} c_{k,t}$.

We now prove (*) by induction on the clique size t. If $t = 1$, the statement is obvious, since First-Fit will assign the same color to every vertex of a graph with no edges. Assume that $t > 1$ and that there exists an on-line algorithm $A_{k,t-1}$ and a constant $c_{k,t-1}$ such that $A_{k,t-1}$ colors any on-line presentation $G^<$ of $G \in q$ Forb(T_k) satisfying $\omega(G) \le t - 1$, with at most $c_{k,t-1}$ colors. To prove the induction step, we must show that there exists an algorithm $A_{k,t}$ and a constant $c_{k,t}$ such that $A_{k,t}$ colors any on-line presentation $G^<$ of $G \in q$ Forb(T_k) satisfying $\omega(G) \le t$, with at most $c_{k,t}$ colors; to this end, we set up a secondary induction. To state the secondary induction, we must refer to three sequences of parameters $p_2, \dots, p_t$, $\rho_2, \dots, \rho_t$, and $w_2, \dots, w_t$. We delay the calculation of these sequences until after a sketch of the secondary induction.

We call a template with s parts and $p_s + k^4$ vertices in each part an s-template. To appreciate the following statement, which we will prove by induction on s, it is important to realize that $p_s + k^4$ will be much smaller than w_s.

(**) For $2 \le s \le t + 1$, there exists an algorithm $A_{k,t,s}$ and a constant $c_{k,t,s}$ such that

(i) For $2 \le s \le t$, if $G^<$ is an on-line presentation of $G \in q$ Forb$(T_k, K_{w_s}^s)$ with $\omega(G) \le t$, then $A_{k,t,s}$ colors $G^<$ using at most $c_{k,t,s}$ colors,

(ii) For $3 \le s \le t + 1$, if $G^<$ is an on-line presentation of a graph $G \in q$ Forb(T_k), where $\omega(G) \le t$ and no $(s - 1)$-template of G has ρ_{s-1} extra points, then $A_{k,t,s}$ colors $G^<$ with at most $c_{k,t,s}$ colors.

Some general comments are in order now. The first comment is that the base step, the case where $s = 2$, follows from Theorem 1.4, regardless of the value of w_2. Note that (ii) makes no assertion in this case. By far, the hardest part of proving (**) is showing that, if (i) is true for $s - 1$, then (ii) is true for s. Next, the sequences of parameters will be defined in such a way that, whenever we have (ii) for a particular value of s, $3 \le s \le t$, we obtain (i) for the same value of s as an immediate corollary. Finally, we will define $\rho_t = 1$, so that (ii) in the case where $s = t + 1$ implies that we may prove the primary induction by putting $A_{k,t} = A_{k,t,t+1}$ and $c_{k,t} = c_{k,t,t+1}$. If $\omega(G) \le t$, a t-template of G cannot have any extra points, since no vertex of G can have neighbors in every part of a t-template.

We now state the properties that our sequences of parameters must have for us to prove (**).

(a′) If $G \in q$ Forb$(K_{w_s}^s)$ and B is an $(s - 1)$-template of G, then B has less than ρ_{s-1} extra points, for $3 \le s \le t$.

(b′) If G is a graph and B is an s-template of G that has k extra points, $x_1, \dots, x_k$, then $N(x_1) \cap \cdots \cap N(x_k) \cap B \ne \emptyset$, for $2 \le s \le t$.

Property (a$'$) is all we need to show that establishing (ii) for a particular value of s yields a proof of (i) for that same value. Suppose that we have defined our parameters so that (a$'$) holds and assume (ii) for some $s \le t$. Let $G^<$ be an on-line presentation of $G \in q\ \mathrm{Forb}(T_k,\ K^s_{w_s})$ with $\omega(G) \le t$. Assuming (ii), we know that, if $A_{k,t,s}$ uses more than $c_{k,t,s}$ colors on $G^<$, then some $(s-1)$-template of G has ρ_{s-1} extra points. However, by (a$'$), $G \notin q\ \mathrm{Forb}(K^s_{w_s})$, a contradiction. Property (b$'$) is used to handle a small technicality that arises when we show that, if (i) is true for $s-1$, then (ii) is true for s. A more detailed motivation for (b$'$) outside the context of the algorithm is impractical.

We now define our parameters, using a "reversed" induction. Let the function w be defined by $w(k, p, \rho) = (p + k^4)(1 + f_1(0, k, \rho) + f_1(1, k, \rho))$ for all positive integers k, p, and ρ.

Base: Let $p_t = k^5$. Let $\rho_t = 1$. Let $w_t = w(k, p_t, \rho_t)$.

Induction: Suppose that p_s, ρ_s, and w_s have been defined for $s > 2$. Then $p_{s-1} = \max\{k^4 w_s, k^5\}$, $\rho_{s-1} = w_s$, and $w_{s-1} = w(k, p_{s-1}, \rho_{s-1})$.

As is the case for property (b$'$), the definition of the function w is motivated by technicalities that arise in a detailed discussion of the algorithm.

We now verify (a$'$). Suppose that $G \in q\ \mathrm{Forb}(K^s_{w_s})$ and B is a template of G with $s-1$ parts, $p_{s-1} + k^4$ vertices in each part, where $3 \le s \le t$. Suppose for contradiction that B has $r = \rho_{s-1} = w_s$ extra points, $x_1, \dots, x_r$. Consider any fixed part Q of B. By the definition of an extra point, x_1 has $p_{s-1} \ge k^4 w_s$ neighbors in Q. At least $k^4 w_s - k^4$ of these neighbors are neighbors of x_2, again by the definition of an extra point. At least $k^4 w_s - 2k^4$ of these points are neighbors of x_3. Continuing in this manner, we may find a subset of Q of size at least $k^4 w_s - ((\rho_{s-1} - 1)k^4) = w_s$, which is adjacent to all of $\{x_1, \dots, x_r\}$. Since Q was chosen arbitrarily, we may do the same in every part of B. This results in an augmented $K^s_{w_s}$, a contradiction.

Property (b$'$) is verified in a similar manner. Suppose that G is a graph, that B is a template of G with s parts, $p_s + k^4$ vertices in each part, and that B has k extra points, $x_1, \dots, x_k$. Then x_1 has at least $p_s \ge k^5 = kk^4$ neighbors in every part of B, at least $kk^4 - k^4$ of which are also neighbors of x_2, and so on. We then find a common neighbor for $x_1, \dots, x_k$. (In fact, we find at least $kk^4 - (k-1)k^4 = k^4$ neighbors in each part of B.)

By our earlier remarks, to prove the induction step of (*), it suffices to prove (**). We have already noted that, in the base step of (**), (i) is a consequence of Theorem 1.4 and (ii) is trivial. By our remarks on property (a$'$), to show the induction step of (**), it suffices to show that, whenever (i) holds for $s-1$, (ii) holds for s, $2 < s \le t+1$. By the primary induction hypothesis, there exists an algorithm $A_{k,t-1}$ and a constant $c_{k,t-1}$ such that, if $G^<$ is. an on-line presentation of a graph $G \in q\ \mathrm{Forb}(T_k)$ with $\omega(G) \le t-1$, then $A_{k,t-1}$ colors $G^<$ with at most $c_{k,t-1}$ colors. By the secondary induction hypothesis, there exists an algorithm $A_{k,t,s-1}$ and a constant $c_{k,t,s-1}$ such that, if $G^<$ is an on-line presentation of $G \in q\ \mathrm{Forb}(T_k,\ K^{s-1}_{w_{s-1}})$ with $\omega(G) \le t$, then $A_{k,t,s-1}$ colors G with at most $c_{k,t,s-1}$ colors. It remains for us to show that, given these hypotheses, there exists an on-line algorithm $A_{k,t,s}$ and a constant $c_{k,t,s}$ satisfying (ii).

For the remainder of the proof, let $p = p_{s-1}$ and $\rho = \rho_{s-1}$. Also, "template" should be read as "$(s-1)$-template." Recall that, if B is a template, then we call a vertex x a strong 1-neighbor of B if x has k neighbors in some part of B. We call x an extra point of B if x has less than k^4 nonneighbors in every part of B. An acceptable template sequence is a template sequence $B_1, \dots, B_r$ such that, if $1 \le m < n \le r$ and $x \in B_n$, then x is not a strong 1-neighbor of B_m.

We now present the algorithm $A_{k,t,s}$. To show (ii), we assume that no template of the graph being presented has ρ_{s-1} extra points. The key feature of the algorithm is that it maintains an acceptable template sequence $B_1, \ldots, B_r$. Whenever a template B_i is added to the sequence, we arbitrarily assign labels $\{1, \ldots, |B_i|\}$ to the vertices of B_i. These labels are not part of the coloring, since a vertex may not become part of a template until long after it has entered the graph. To each template B_i, we will also associate a set of (not necessarily all) strong 1-neighbors, O_i. A vertex x may be assigned to O_i in one of two ways, either when x enters the graph (if B_i had already been formed) or when B_i is formed (if the template doesn't appear until after x has entered). In the latter case, we give x a "shadow" color, as we detail below. As with the labels on the template points, the shadow colors are not part of the coloring produced by the algorithm, but are instead records to be used internally by the algorithm. In either case, the assignment of x to O_i is irrevocable, and the O_i's are pairwise disjoint.

Suppose that when a vertex x enters the graph the acceptable sequence of templates is $B_1, \ldots, B_r$. The algorithm colors x and updates the template list as follows.

Case 1. If x is a strong 1-neighbor of some template in the sequence, find the smallest i such that x is a strong 1-neighbor of B_i. Add x to O_i. Assign x a several coordinate color. The first coordinate identifies x as a vertex that was classified as a strong 1-neighbor at the time it entered the graph. The second coordinate is the set of labels used on $N(x) \cap B_i$. Note that there are a fixed number of labels since all the templates have the same size. To compute the third coordinate, apply the algorithm $A_{k,t-1}$ (which exists by the primary induction hypothesis) to the subgraph induced by vertices of O_i that received the same colors as x in their first two coordinates. Since these vertices have a common neighbor in B_i, the induction hypothesis implies that we use at most $c_{k,t-1}$ colors in this coordinate. Let V' be the set of vertices that received the same color as x in their first three coordinates. Note that $V' \cap O_j$ is an independent set for $j = 1, \ldots, r$.

Claim A. The subgraph induced by V' and the independent sets $V' \cap O_j$ satisfy conditions (a) and (b) of Lemma 5.2 with $d = (k + \rho)f_1(1, k, \rho)$, $e = k + \rho$, and $f = (s - 1)(p + k^4)f_1(1, k, \rho)f_1(2, k, \rho)$.

Given this claim, by Lemma 5.2, we may apply an on-line algorithm to the subgraph induced by vertices that received the same color as x in their first three coordinates. The number of colors used in each coordinate will be bounded in terms of k, ρ, s, and $\omega(G)$, all of which are bounded in terms of t, and it will be a proper coloring.

Case 2. If x is not a strong 1-neighbor for any template in the sequence at the time x enters, then we attempt to find a set of vertices that, together with x, form a template that may be added to the sequence without violating the key properties of the sequence. That is, we look for a set B_{r+1} such that B_{r+1} is an $(s - 1)$-template and $B_{r+1} \cap (B_i \cup O_i) = \emptyset$ for $1 \leq i \leq r$. Note that, if B_{r+1} satisfies this condition, no vertex of B_{r+1} is a strong 1-neighbor of any earlier template in the sequence. If such a set B_{r+1} can be found, add B_{r+1} to the template sequence and assign labels to the vertices of B_{r+1}. Assign x a two-coordinate color. The first coordinate identifies x as a vertex that was used to form a new template at the time it entered. The second coordinate is computed by applying First-Fit to the subgraph induced by vertices that received the same color as x in their first coordinate. Note that every template in the acceptable sequence contains precisely one such vertex. Since we use First-Fit in the second coordinate, we know that x will be properly colored. Moreover, it is easy to see that the algorithm will use a bounded number of colors in the second coordinate for the subgraph induced by vertices that were used to form templates at the time they entered. This is because, by Lemma 4.1 and the fact that x is not an extra point of any previous template (it is not even a strong 1-neighbor), this subgraph has degree at most $f_1(0, k, \rho)$.

If y is a vertex that entered before x such that, for $1 \leq i < r$, $y \notin B_i \cup O_i$ (with y a strong 1-neighbor of B_{r+1}), then assign y to O_{r+1}. The algorithm then assigns to y a "shadow" color $c(y)$; this color is strictly for record-keeping purposes, since y's "real" color was assigned when y entered. The shadow color is assigned as follows. Imagine that a "twin" vertex y' is presented immediately after x and that y' has precisely the same neighbors as y (at the time x enters and thereafter). Apply the algorithm to y' as if it were an actual point presented and, for any points in the graph that have already received a shadow color, use the shadow color, rather than the color actually assigned, to compute the color for y'. In fact, y' would be colored under Case 1, because y' looks exactly like y, which is now a strong 1-neighbor of B_{r+1}. When other vertices require shadow colors in the future, y' will be treated as if it had been actually presented. This will guarantee that the shadow colors form a proper coloring of the set of vertices that received shadow colors. If there is more than one such y, say $y_1, \ldots, y_m$, when B_{r+1} is formed, then apply the same procedure to each vertex. Finally, the algorithm assigns shadow colors to all the vertices of B_{r+1}, except x.

Claim B. If x is a vertex with shadow color $c(x)$, then $c(x) \notin S(x) = \{c(y): y \in N(x)\}$.

Case 3. If x is not a strong 1-neighbor of any template at the time x entered and if x cannot be used to form a new template for the sequence, then we assign x a three-coordinate color as follows. The first coordinate identifies x as a vertex that could not be colored in either Case 1 or Case 2. The second coordinate of x is the set $S(x) = \{c(y): y \in N(x)\}$. Note that the number of colors used in the second coordinate is 2^b, where b is the maximum number of colors used in step 1; assuming Claim A, b is bounded in terms of t, so 2^b is, as well. To compute the third coordinate of x, apply the algorithm $A_{k,t,s-1}$ (whose existence is asserted by the secondary induction hypothesis) to the subgraph G_S induced by vertices that received the same colors as x in their first two coordinates, where $S = S(x)$ is the second coordinate of x's color.

Claim C. The subgraph G_S does not contain an augmented $K_{w_{s-1}}^{s-1}$.

Assuming Claim C, by the secondary induction hypothesis, $A_{k,t,s-1}$ gives a proper coloring of G_S, and therefore x is properly colored. Moreover, the secondary induction hypothesis implies that no more than $c_{k,t,s-1}$ colors are used in the third coordinate.

By the remarks included in the statement of the algorithm, to finish the proof of (**), it suffices to prove Claims A–C.

Proof of Claim A. Since it is the simpler of the two, we first verify that condition (b) of Lemma 5.2 is satisfied. That is, we check that, whenever V' contains an augmented $K_{d,d}$, say H, there exist integers α and β such that $H \cap (O_\alpha \cup O_\beta)$ contains an augmented $K_{e,e}$, where $d = (k + \rho)f_1(1, k, \rho)$ and $e = k + \rho$. Consider a complete bipartite graph H in V' with $(k + \rho)f_1(1, k, \rho)$ vertices in each part. Let H_1 and H_2 be the independent sets of H. If $y \in H_1$, then $y \in O_j$, for exactly one j, $1 \leq j \leq r$. Note that, if $I = \{j : \exists y \in O_j \cap H_1\}$, $|I| < f_1(1, k, \rho)$. To see this, suppose without loss of generality that $\{1, \ldots, f_1(1, k, \rho)\}$ is contained in I. Then, since no template has ρ extra points and $|H_1| > \rho f_1(1, k, \rho)$, by the Pigeonhole Principle, some vertex y' of H_2 is not an extra point of templates $B_1, \ldots, B_g$, where $g = f_1(1, k, \rho)$. Then, however, we have a contradiction of Lemma 4.1, since y' is adjacent to all of H_1. Because of this bound on $|I|$, using the Pigeonhole Principle, we may find an index α such that $|O_\alpha \cup H_1| = k + \rho$. By a similar argument, we find a subset of $H_2 \cap O_\beta$ of size $k + \rho$. This completes the verification of condition (b).

We now verify (a) of Lemma 5.2; that is, for all α, $1 \leq \alpha \leq r$, $|\{\beta: (V' \cap O_\alpha) \cup (V' \cap O_\beta)$ contains an augmented $K_{e,e}\}| \leq f$, where $e = k + \rho$ and $f = (s - 1) \times (p + k^4)f_1(1, k, \rho)f_1(2, k, \rho)$. Consider an arbitrary node B_α of the auxiliary graph (or, more precisely, the node of the auxiliary graph corresponding to the template B_α). We wish to show first that, for every edge in the auxiliary graph that is incident on B_α (say the other endpoint is B_β), there exist a vertex $x_\beta \in B_\alpha$ and a strong 1-neighbor y_β of B_α such that

(i) y_β is at the top level of a 2-tube with 0-level B_β,

(ii) x_β and y_β are not extra points of B_β,

(iii) $x_\beta \sim y_\beta$.

This is illustrated in Fig. 6.1.

First, note that, if there exists a complete bipartite graph with $k + \rho$ vertices in each part where one part, say X, is contained in $V' \cap O_\alpha$ and the other, say Y, in $V' \cap O_\beta$ for some $\alpha \neq \beta$, then there exists a 2-tube such that its 0-level is B_β, its first level is contained in Y (choose any k vertices of Y), and the vertex y_β at the top level of the tube is an element of X that is not an extra point of B_β. Note that y_β exists because some vertex of X must fail to be an extra point of B_β, since $|X| > \rho$. Moreover, since every vertex of X is a strong 1-neighbor of B_α, we may find k vertices in one part of B_α that are neighbors of y_β. One of these k vertices must fail to be an extra point of B_α: if $\alpha > \beta$, then no element of B_α is an extra point (or even a strong 1-neighbor) of B_β. If $\beta > \alpha$ and k vertices in one part of B_α are extra points of B_β, then they have a common neighbor z in B_β, by property (b'). Then, however, z is a strong 1-neighbor of B_α, and this fact would have prevented us from adding B_α to the template sequence. Thus we may choose x_β to be any neighbor of y_β that is not an extra point of B_β.

Thus, if the degree of B_α is $(s - 1)(p + k^4)f_1(1, k, \rho)f_1(2, k, \rho)$ or higher, we find, by the Pigeonhole Principle (there are only $(s - 1)(p + k^4)$ vertices in B_α), a vertex x of B_α that has a neighbor in the top level of $f_1(1, k, \rho)f_1(2, k, \rho)$ 2-tubes with distinct 0-levels. (That is, $x = x_\beta$ for $f_1(1, k, \rho)f_1(2, k, \rho)$ distinct values of β.) See Fig. 6.2. Since the algorithm forces the O_i's to be pairwise disjoint, the 1-levels of these tubes are pairwise disjoint. It is still possible, however, that the 2-levels may not be disjoint. For each β, however, we choose the y_β so that it will not be extra point of B_β. Thus no vertex $y = y_\beta$ can be at the top of $f_1(1, k, \rho)$ of these tubes, or else there exist $f_1(1, k, \rho)$ disjoint 1-tubes whose bases form an acceptable template sequence and whose top levels are all adjacent to y, contradicting Lemma 4.1. Thus we may find among the collection of 2-tubes whose top points are adjacent to x a subcollection of $f_1(2, k, \rho)$ pairwise disjoint 2-tubes. When these tubes are ordered so that their bases are a subsequence of the ac-

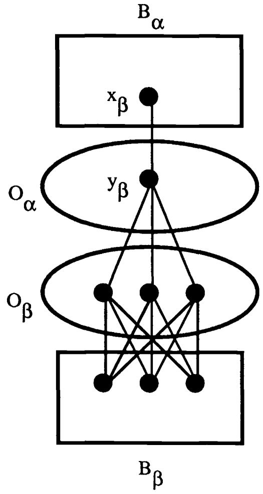


FIG. 6.1.

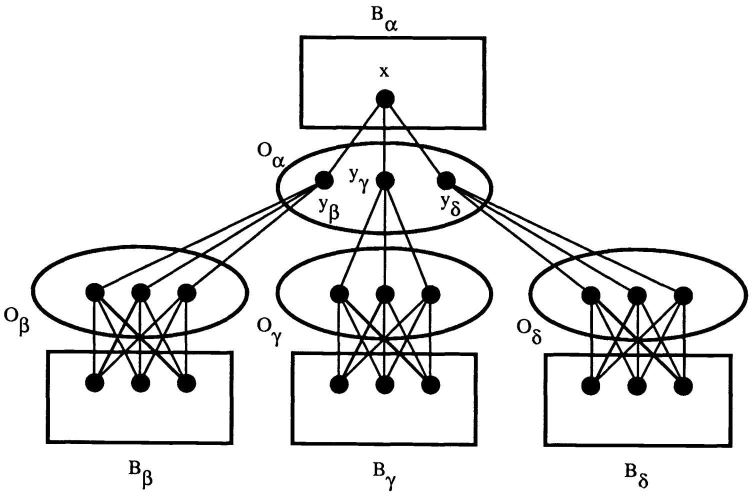


FIG. 6.2.

ceptable template sequence generated by the algorithm, the hypotheses of Lemma 4.1 are satisfied, so this is a contradiction. This shows that (a) of Lemma 5.2, i.e., the degree condition for the auxiliary graph, holds as claimed. □

Proof of Claim B. If $c(x)$ is the shadow color of x and $c(x) \in S(x)$, then there exists some vertex y with shadow color $c(y) = c(x)$ such that y received its shadow color before x entered the graph and $y \sim x$. However, since $y \sim x$ and y had a shadow color at the time the shadow color of x was assigned, $c(y) \neq c(x)$. □

Proof of Claim C. We argue by contradiction. Suppose that, at some point in the algorithm, some G_S contains an augmented $K^{s-1}_{w_{s-1}}$. Let Q be this subgraph and let $Q_1, \ldots, Q_{s-1}$ be the parts of Q. Let x be the last vertex of Q to enter the graph. Without loss of generality, $x \in Q_{s-1}$. Let $B_1, \ldots, B_r$ be the templates in the sequence at the time x entered.

We first claim that, for $1 \leq j \leq s-2$, no element of Q_j has been added to any B_j or O_j before x entered; if there were such a vertex, say y, then y would have received a shadow color $c(y)$. Then, however, $c(y) \in S(x) = S = S(y)$, contradicting Claim B.

Now consider Q_{s-1}. Let z be the last vertex of $Q' = Q_1 \cup \cdots \cup Q_{s-1}$ to enter the graph. Note that, by the above argument, z is not an extra point of any template in the sequence at the time x entered. (Indeed, this is true of all the vertices in Q'.) Label the vertices of Q_{s-1} as follows. For each $y \in Q_{s-1}$, assign y, if possible, the smallest m such that, at the time x entered, either $y \in B_m$ or $y \in O_m$. If no such m exists, assign y the label ∞. Note that the fact that z is not an extra point of any template in the sequence at the time x entered and the fact that z is adjacent to all of Q_{s-1} imply that at most $1 + f_1(0, k, \rho) + f_1(1, k, \rho)$ labels are used. Thus we find a set of vertices of size $p + k^4$ all of whose elements have the same label. In this case, however, ∞ cannot be the common label, or else the algorithm would have been able to add some template from Q to the sequence (and thereby give some vertex in Q a color according to Case 2). Thus the common label is a natural number m. Let $x_1, \ldots, x_a$, where $a = p + k^4$, be the vertices of Q_{s-1} that received the label m. Each of $x_1, \ldots, x_a$ must have entered the graph before B_m was formed; any that entered after B_m was formed would have been a strong 1-neighbor at the time it entered, and thus would have been colored in Case 1. On the other hand, z must have entered after B_m was formed: by the manner in which sets of strong 1-neighbors are formed, $x_1, \ldots, x_a$ are not strong 1-neighbors of any template preceding B_m in the sequence; else, they would have been assigned to, say, O_n where $n < m$, as soon as B_n was formed. Hence, if z, and therefore all of Q', entered before B_m was formed, the algorithm would have added a template to the sequence and the last of the points $z, x_1, \ldots, x_a$ to enter the graph would have been colored by Case 2. It cannot be the case that z entered at the time B_m was formed; else, z would have been colored in Case 2. If, however, z enters G after B_m is formed, then, x_1, for example, had received a shadow color by the time z entered, so $c(x_1) \in S(z)$. Then $c(x_1) \in S(z) - S(x_1)$ and $S(z) \neq S(x_1)$, a contradiction. □

This completes the induction step of (**). The induction step of (*) follows, and the proof of the theorem is complete. □

REFERENCES

[B] D. BEAN, *Effective coloration*, J. Symbolic Logic, 41 (1976), pp. 469–480.

[C] V. CHVÁTAL, *Perfectly ordered graphs*, in Topics on Perfect Graphs, Ann. Discrete Math., 21 (1989), pp. 63–65.

[EH] P. ERDÖS AND A. HAJNAL, *On chromatic number of graphs and set systems*, Acta Math. Acad. Sci. Hungar., 17 (1966), pp. 61–99.

[G1] A. GYÁRFÁS, *On Ramsey Covering-Numbers*, Coll. Math. Soc. János Bolyai 10, Infinite and Finite Sets, North-Holland/American Elsevier, New York, 1975, pp. 801–816.

[G2] ———, *Problems from the world surrounding perfect graphs*, Zastosowania Matematyki Applicationes Mathemacticae, 19 (1985), pp. 413–441.

[Ga] T. GALLAI, *Transitiv orientierbare Graphen*, Acta. Math. Acad. Sci. Hungar., 18 (1967), pp. 25–66.

[GL1] A. GYÁRFÁS AND J. LEHEL, *On-line and first-fit coloring of graphs*, J. Graph Theory, 12 (1988), pp. 217–227.

[GL2] ———, *First-fit and on-line chromatic number of families of graphs*, Ars Combinatoria, 29C (1990), pp. 168–176.

[GL3] ———, *Effective on-line coloring of P_5-free graphs*, Combinatorica, 11 (1991), pp. 181–184.

[GRS] R. GRAHAM, B. ROTHSCHILD, AND J. SPENCER, *Ramsey Theory*, John Wiley, New York, 1980.

[GST] A. GYÁRFÁS, E. SZEMERÉDI, AND TUZA, *Induced subtrees in graphs of large chromatic number*, Discrete Math., 30 (1980), pp. 235–244.

[K1] H. A. KIERSTEAD, *An effective version of Dilworth's theorem*, Trans. Amer. Math., Soc., 268 (1981), pp. 63–77.

[K2] ———, *Recursive colorings of highly recursive graphs*, Canad. J. Math., 33 (1981), pp. 1279–1290.

[K3] ———, *An effective version of Hall's theorem*, Proc. Amer. Math. Soc., 88 (1983), pp. 124–128.

[K4] ———, *The linearity of First-Fit for coloring interval graphs*, SIAM J. Discrete Math., 1 (1988), pp. 526–530.

[KP1] H. A. KIERSTEAD AND S. G. PENRICE, *Radius two trees specify χ-bounded classes*, J. Graph Theory, to appear.

[KP2] ———, *Recent results on a conjecture of Gyárfás*, Congr. Numer., 79 (1990), pp. 182–186.

[KPT] H. A. KIERSTEAD, S. G. PENRICE, AND W. T. TROTTER, *On-line and first-fit colorings of graphs which do not induce P_5*, SIAM J. Discrete Math., submitted.

[KR] H. A. KIERSTEAD AND V. RÖDL, *Applications of Hypergraph Coloring to Coloring Graphs Which Do Not Induce Certain Trees*, preprint.

[KT1] H. A. KIERSTEAD AND W. T. TROTTER, *An extremal problem in recursive combinatorics*, Congr. Numer., 33 (1981), pp. 143–153.

[KT2] ———, *Colorful induced subgraphs*, Discrete Math., 101 (1992), pp. 165–169.

[S1] J. H. SCHMERL, *Recursive colorings of graphs*, Canad. J. Math., 32 (1980), pp. 821–831.

[S2] ———, *The effective version of Brooks' theorem*, Canad. J. Math., 34 (1982), pp. 1036–1046.

[Su] D. P. SUMNER, *Subtrees of a graph and chromatic number*, in The Theory and Applications of Graphs, G. Chartrand, ed., John Wiley, New York, 1981, pp. 557–576.

[W] D. R. WOODALL, *Problem No. 4, Combinatorics*, in Proc. British Combinatorial Conference, 1973, London Math. Soc. Lecture Note Series 13, T. P. McDonough and V. C. Marvon, eds., Cambridge University Press, Cambridge, UK, 1974, p. 202.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 90–101, February 1994

009

LITTLEWOOD-OFFORD INEQUALITIES FOR RANDOM VARIABLES*

I. LEADER† AND A. J. RADCLIFFE‡

Abstract. The *concentration* of a real-valued random variable X is

$$c(X) = \sup_{t \in \mathbf{R}} \mathbf{P}(t < X < t + 1).$$

Given bounds on the concentrations of n independent random variables, how large can the concentration of their sum be?

The main aim of this paper is to give a best possible upper bound for the concentration of the sum of n independent random variables, each of concentration at most $1/k$, where k is an integer. Other bounds on the concentration are also discussed, as well as the case of vector-valued random variables.

Key words. Littlewood–Offord problem, concentration, normed spaces

AMS subject classifications. 60G50, 06A07, 52A40

Introduction. In 1943, Littlewood and Offord [8], concerned with estimating the number of real zeros of random polynomials, proved that, given complex numbers $(a_i)_1^n$ of modulus at least 1, not too many of the sums $s_A = \sum_{i \in A} a_i$, $A \subset \{1, 2, \dots, n\}$ lie in any open disc of diameter 1. They showed that the maximum number is $O(2^n n^{-1/2} \log n)$.

In 1945, Erdös [2] noted that, if the a_i are real numbers, then Sperner's theorem—on the maximum size of an antichain in the poset $\mathcal{P}(n) = \mathcal{P}(\{1, 2, \dots, n\})$—implies a best possible upper bound. Indeed, suppose first that the a_i are all positive. Then, given an open interval I of length 1, the set system $\mathcal{A}_I = \{A \subset \{1, 2, \dots, n\} : s_A \in I\}$ is an antichain, since, if $B \supset A$, then $s_B - s_A = s_{B \setminus A} \geq |B \setminus A| \geq 1$. Thus, for all I, $\mathcal{A}_I \leq \binom{n}{\lfloor n/2 \rfloor}$ by Sperner's theorem [9]. The result for positive reals immediately implies that the same conclusion follows for all reals. Kleitman [5] and Katona [4] independently showed that the same bound, of $\binom{n}{\lfloor n/2 \rfloor}$, holds for $(a_i)_1^n$ in $\mathbb{C}$, thus giving a best possible improvement of the lemma of Littlewood and Offord. In [6] Kleitman proved a considerable extension of this result, namely, to sums of vectors $(a_i)_1^n$ of norm at least 1 in an arbitrary normed space, thus setting a conjecture of Erdös.

Jones [3] suggested a probabilistic framework for these questions, regarding a vector $a \neq 0$ in a normed space E as being naturally associated with an E-valued random variable X_a with $\mathbf{P}(X_a = 0) = \frac{1}{2}$ and $\mathbf{P}(X_a = a) = \frac{1}{2}$. So, if δ_a is the delta measure on E concentrated at a, then the distribution of X_a is $\frac{1}{2}(\delta_0 + \delta_a)$. Kleitman's result can then be stated as follows.

THEOREM A (see [6]). *Let $(a_i)_1^n$ be vectors in a normed space E of norm at least 1 and let $(X_i)_1^n$ be independent random variables with X_i having distribution $\frac{1}{2}(\delta_0 + \delta_{a_i})$. Then, for any open set $U \subset X$ of diameter at most 1, we have*

$$\mathbf{P}\left(\sum_1^n X_i \in U\right) \leq 2^{-n} \binom{n}{\lfloor n/2 \rfloor}.$$

Note that this bound is clearly best possible, equality being attained if, for instance, all the a_i are equal.

* Received by the editors November 11, 1991; accepted for publication (in revised form) November 4, 1992.

† Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England.

‡ Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213-3890.

The conclusion of Theorem A gives a bound on the extent to which the values of the random variable $\sum X_i$ are concentrated in one place. This prompts the following definition.

Let E be a normed space. The *concentration* of an E-valued random variable X is $c(X) = \sup \mathbf{P}(X \in U)$, where the supremum is taken over all open subjects $U \subset E$ having diameter at most 1.

The hypotheses of Theorem A can also be stated in terms of concentration, and in this form it reads as follows.

THEOREM A′. *Let $(X_i)_1^n$ be independent E-valued random variables that are essentially two-valued and have concentration at most $\frac{1}{2}$. Then*

$$c\left(\sum_1^n X_i\right) \leq 2^{-n}\binom{n}{\lfloor n/2 \rfloor}.$$

The main result of this paper is a result that extends Theorem A′ in the case when $E = \mathbb{R}$ by removing the restriction that each X_i be essentially two-valued.

THEOREM 1. *Let $(X_i)_1^n$ be independent real-valued random variables of concentration at most $\frac{1}{2}$. Then*

$$c\left(\sum_1^n X_i\right) \leq 2^{-n}\binom{n}{\lfloor n/2 \rfloor}.$$

Our technique is closely related to that of Kleitman, being based on symmetric chain decompositions. In §1 we give a proof of Theorem 1, as well as presenting some background about symmetric chain decompositions.

In §2 we consider sums of random variables of concentration at most $1/q$, where q is an integer, and we generalise some results of Jones [3]. To state our result, we need some fairly standard notation. We write $[q]$ for the set $\{0, 1, \ldots, q-1\}$ and also for the poset with that ground set and the natural ordering. We write $[q]^n$ for the product of n copies of $[q]$ with the usual product ordering, i.e., $(x_i)_1^n \leq (y_i)_1^n$ if and only if $x_i \leq y_i$ for each i. Finally, we write W for the size of the largest level set in the ranked poset $[q]^n$ as follows:

$$W = W_{q,n} = |\{(x_i)_1^n \in [q]^n : \sum x_i = \lfloor n(q-1)/2 \rfloor\}|.$$

THEOREM 2. *Let $(X_i)_1^n$ be independent real-valued random variables with $c(X_i) \leq 1/q$, where $q \in \mathbb{N}$. Then $c(\sum_1^n X_i) \leq W/q^n$.*

This bound is clearly best possible. Equality is attained when, for instance, each X_i has distribution $(1/q)(\delta_0 + \delta_1 + \cdots + \delta_{q-1})$.

Based on Theorem 2, we are perhaps tempted to guess that the sum of n independent random variables, each of concentration at most p/q (p and q coprime integers), has concentration bounded by the proportion of $[q]^n$ occupied by the largest p layers. Unfortunately, very simple examples show that this is not the case. Rather surprisingly, given this, the result does hold when $p = 2$. Both the examples and the proof are given in §3.

Finally, in §4, we turn our attention to the vector-valued case. We consider some of the problems raised by Jones [3] and answer some of his questions.

1. Sums of random variables of concentration at most $\frac{1}{2}$. Before considering the details of our proof of Theorem 1, some discussion of symmetric chain decompositions is in order. These will prove to be vital for our results, as they were for Kleitman's.

A *symmetric chain decomposition* of the power set $\mathcal{P}(n) = \mathcal{P}\{1, 2, \ldots, n\})$ is a partition of $\mathcal{P}(n)$ into chains (totally ordered subsets) in such a way that each chain

$\mathcal{A} = \{A_1, A_2, \ldots, A_r\}$ with $A_1 \subset A_2 \subset \cdots \subset A_r$ satisfies $|A_{k+1}| = |A_k| + 1$ and $|A_1| + |A_r| = n$. Thus each $\mathcal{A}$ is arrayed symmetrically about the "middle layer" of $\mathcal{P}(n)$ and contains one set from each layer between the extremes. In particular, of course, $\mathcal{A}$ must contain one set of size $\lfloor n/2 \rfloor$. de Bruijn, Tengbergen, and Kruyswijk [1] showed that symmetric chain decompositions do exist, and Kleitman's beautiful result, Theorem A, was based on that proof.

Their proof goes as follows. Suppose that $(\mathcal{A}_j)_1^s$ is a symmetric chain decomposition of $\mathcal{P}(n-1)$. We can construct a symmetric chain decomposition of $\mathcal{P}(n)$ in the following manner. Take a copy of $(\mathcal{A}_j)_1^s$ in each layer of $\mathcal{P}(n)$: the bottom layer, which is exactly $\mathcal{P}(n-1)$, and the top layer (of sets containing n). This is very definitely not a symmetric chain decomposition of $\mathcal{P}(n)$, but, by transferring the top element of each chain in the top layer to the corresponding chain downstairs, everything can be fixed. More precisely, for each chain $\mathcal{A}_j = \{A_1, A_2, \ldots, A_r\}$ with $A_1 \subset A_2 \subset \cdots \subset A_r$ set

$$\mathcal{A}'_j = \{A_1, A_2, \ldots, A_r, A_r \cup \{n\}\},$$

$$\mathcal{A}''_j = \{A_1 \cup \{n\}, A_2 \cup \{n\}, \ldots, A_{r-1} \cup \{n\}\}.$$

The collection $\{\mathcal{A}'_j, \mathcal{A}''_j\}_1^s$ forms a symmetric chain decomposition of $\mathcal{P}(n)$, after the removal of those $\mathcal{A}''_j$ that are empty.

A sequence $(m_j)_1^s$ is called a *symmetric profile* for $\mathcal{P}(n)$ if, for some (and therefore, up to rearrangement, every) symmetric chain decomposition of $\mathcal{P}(n)$, say $(\mathcal{A}_j)_1^s$, we have $m_j = |\mathcal{A}_j|$. Note that $s = \binom{n}{\lfloor n/2 \rfloor}$.

As the above proof shows, we get a symmetric profile for $\mathcal{P}(n+1)$ by taking $(m_i)_1^s$ and replacing each m_j by the pair $m_j - 1$, $m_j + 1$ and then discarding zeros. At first, this may seem not to the point, since we can write the symmetric profile for $\mathcal{P}(n)$ easily and explicitly: A sequence $(m_j)_1^s$ is a symmetric profile for $\mathcal{P}(n)$ if $s = \binom{n}{\lfloor n/2 \rfloor}$ and the number of j with $m_j = n + 1 - 2i$ is $\binom{n}{i} - \binom{n}{i-1}$ (with the convention that $\binom{n}{-1} = 0$). However, symmetric chain decompositions will arise in more complicated situations, in which finding an explicit expression is much harder. Fortunately, all that we need for the proofs is the total number of chains in a decomposition and the way in which the symmetric profile changes as the poset grows. To illustrate this, below is Kleitman's proof of Theorem A, using symmetric profiles.

Proof of Theorem A. The values of $X = \sum_1^n X_i$ are exactly those vectors in E of the form $x_A = \sum_{i \in A} a_i$, where A is any subset of $\{1, 2, \ldots, n\}$. The distribution of X is $2^{-n} \sum_{A \subset \{1,2,\ldots,n\}} \delta_{x_A}$. To show that $c(X)$ is small, we partition $\mathcal{P}(n)$ into subsets $(\mathcal{A}_j)_1^s$ with $s = \binom{n}{\lfloor n/2 \rfloor}$ and

$$(*) \qquad A, B \in \mathcal{A}_j \Rightarrow \|x_A - x_B\| \geq 1.$$

To do this, we in fact do more, namely, prove that the partition can be chosen with $(|\mathcal{A}_j|)_1^s$ being a symmetric profile for $\mathcal{P}(n)$. Once this is proved, the theorem follows easily, since

$$\begin{aligned} \mathbf{P}(X \in U) &= 2^{-n} \sum_{A \subset \{1,2,\ldots,n\}} \delta_{x_A}(U) \\ &= 2^{-n} \sum_{j=1}^{s} \sum_{A \in \mathcal{A}_j} \delta_{x_A}(U) \\ &\leq 2^{-n} s \\ &= 2^{-n} \binom{n}{\lfloor n/2 \rfloor}. \end{aligned}$$

The last inequality holds, since, by $(*)$, at most one x_A with $A \in \mathcal{A}_j$ belongs to U.

The proof goes by induction on n. The result is trivial for $n = 1$, so we turn to the induction step. Take an appropriate partition $\mathscr{P}(n-1) = \bigcup_1^{s'} \mathscr{A}_j$ (where $s' = \binom{n-1}{\lfloor n-1/2 \rfloor}$). Take a support functional $f \in X^*$ for a_n, a functional with $\|f\| = 1$ and $f(a_n) = \|a_n\| \geq 1$. For any $\mathscr{A}_j = \{A_1, A_2, \dots, A_r\}$, choose l with

$$f(x_{A_l}) \geq f(x_{A_k}), \qquad k = 1, 2, \dots, r$$

and set

$$\mathscr{A}_j' = \{A_1, A_2, \dots, A_r, A_l \cup \{n\}\},$$

$$\mathscr{A}_j'' = \{A_1 \cup \{n\}, A_2 \cup \{n\}, \dots, A_{l-1} \cup \{n\}, A_{l+1} \cup \{n\}, \dots, A_r \cup \{n\}\}.$$

The partition of $\mathscr{P}(n)$ that is needed consists of all the nonempty $\mathscr{A}_j'$ and $\mathscr{A}_j''$. Clearly, each $\mathscr{A}_j''$ satisfies $(*)$, since $\mathscr{A}_j$ did originally. In $\mathscr{A}_j'$, we need only check that, for each $A_k \in \mathscr{A}_j$, the norm of $x_{A_l \cup \{n\}} - x_{A_k}$ is large. This follows by applying f as follows:

$$\begin{aligned} \|x_{A_l \cup \{n\}} - x_{A_k}\| &\geq f(x_{A_l \cup \{n\}} - x_{A_k}) \\ &= f(a_n) + f(x_{A_l}) - f(x_{A_k}) \\ &\geq f(a_n) \geq 1. \end{aligned}$$

The profile of the new partition is a symmetric profile of $\mathscr{P}(n)$, since each $\mathscr{A}_j$ of size m splits into two, of sizes $m+1$ and $m-1$. Thus by induction the theorem is proved. □

In the proof of Theorem 1, we will be dealing with random variables and their distributions, treating the latter similarly to the finite subsets of E that arise in the proof of Theorem A. Indeed, we often regard a finite subset of $\mathbb{R}$ as corresponding to a random variable that assigns equal mass to those points and none to all others. We are interested in the distribution of the sum of these random variables, that is, in the convolution of their distributions.

More generally, we will be dealing with finite (positive Borel) measures on $\mathbb{R}$—the collection of all such we denote by $\mathscr{M}$. However, we wish to stress that we will not really be using any measure theory. Indeed, a reader who considers only measures of finite support will not be losing much.

For $\mu \in \mathscr{M}$, the *mass* of μ is $|\mu| = \mu(\mathbb{R})$. Just as before, the *concentration* of μ is $c(\mu) = \sup \mu(I)$, where the supremum is over all open intervals of length 1. The convolution of μ, $\lambda \in \mathscr{M}$ is denoted by $\mu * \lambda$, and we write $\mathscr{M}(m, c)$ for the set of all $\mu \in \mathscr{M}$ of mass m and concentration at most c.

Two elementary facts are summarised in the following lemma.

LEMMA 3. *If μ has mass m and λ has concentration at most c, then $\mu * \lambda$ has concentration at most mc. Also, if $(\mu_i)_1^n \subset \mathscr{M}$, then $c(\sum_1^n \mu_i) \leq \sum_1^n c(\mu_i)$.*

Proof. Given any interval $I = (t, t+1)$, we have

$$\begin{aligned} \mu * \lambda(I) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \chi_I(x+y)\, d\lambda(x)\, d\mu(y) \\ &= \int_{\mathbb{R}} \lambda(I - y)\, d\mu(y) \\ &\leq \int_{\mathbb{R}} c\, d\mu(y) = mc, \end{aligned}$$

proving the first statement. The second is immediate. □

A standard approach in the proofs will be to split up a measure μ into parts with almost disjoint support. We need notation for these parts and therefore we define

$$\text{Left}\,(\mu, m) = \mu|_{(-\infty, t]} - x\delta_t,$$

where $t = \sup \{x : \mu(-\infty, x) \le m\}$ and $x = \mu(-\infty, t] - m$. Thus the support of Left (μ, m) is contained in $(-\infty, t]$ and $|\text{Left}\,(\mu, m)| = m$. We define Right (μ, m) in a similar fashion.

The next lemma, which is rather technical, enables us to peel off, from a convolution of measures of concentration at most 1, a part that also has concentration at most 1. This process is analogous to the transfer that occurs in the de Bruijn / Tengbergen / Kruyswijk proof of the existence of symmetric chain decompositions.

LEMMA 4. *If μ and λ are measures of concentration* 1 *and mass at least* 1, *then, writing μ_{R} for* Right $(\mu, 1)$ *and λ_{L} for* Left $(\lambda, 1)$, *the measure $\nu = \mu_{\mathrm{R}} * \lambda + \mu * \lambda_{\mathrm{L}} - \mu_{\mathrm{R}} * \lambda_{\mathrm{L}}$ belongs to $\mathcal{M}(m(\mu) + m(\lambda) - 1, 1)$.*

Proof. Let $I = (t, t + 1)$ be an interval of length 1 in $\mathbb{R}$. We wish to show that $\nu(I) \le 1$. Set $\mu_{\mathrm{L}} = \mu - \mu_{\mathrm{R}}$ and $x_\mu = \inf \{t : \mu(t, \infty) \le 1\}$. Similarly, let $\lambda_{\mathrm{R}} = \lambda - \lambda_{\mathrm{L}}$ and $x_\lambda = \sup \{t : \lambda(-\infty, t) \le 1\}$. Then we can also write ν as $\mu_{\mathrm{L}} * \lambda_{\mathrm{L}} + \mu_{\mathrm{R}} * \lambda_{\mathrm{L}} + \mu_{\mathrm{R}} * \lambda_{\mathrm{R}}$. If $\mu_{\mathrm{L}} * \lambda_{\mathrm{L}}(I) = 0$, then $\nu(I) = \mu_{\mathrm{R}} * \lambda(I) \le 1$, the last since μ_{R} has mass 1 and λ has concentration 1. Similarly, we are finished if $\mu_{\mathrm{R}} * \lambda_{\mathrm{R}}(I) = 0$. If neither is zero, then necessarily $t \le x_\lambda + x_\mu \le t + 1$. In this case, we split λ yet further. Write

$$\lambda_{\mathrm{LL}} = \lambda_{\mathrm{L}}|_{(-\infty, t - x_\mu]}, \qquad \lambda_{\mathrm{LR}} = \lambda_{\mathrm{L}}|_{(t - x_\mu, \infty)},$$
$$\lambda_{\mathrm{RL}} = \lambda_{\mathrm{R}}|_{(-\infty, t + 1 - x_\mu]}, \qquad \lambda_{\mathrm{RR}} = \lambda_{\mathrm{R}}|_{[t + 1 - x_\mu, \infty)}.$$

Note that $\mu_{\mathrm{R}} * \lambda_{\mathrm{RR}}(I) = \mu_{\mathrm{L}} * \lambda_{\mathrm{LL}}(I) = 0$, so

$$\begin{aligned}\nu(I) &= (\mu_{\mathrm{R}} * \lambda_{\mathrm{LL}} + \mu_{\mathrm{L}} * \lambda_{\mathrm{LR}} + \mu_{\mathrm{R}} * \lambda_{\mathrm{LR}} + \mu_{\mathrm{R}} * \lambda_{\mathrm{RL}})(I) \\ &= (\mu * \lambda_{\mathrm{LR}})(I) + (\mu_{\mathrm{R}} * (\lambda_{\mathrm{LL}} + \lambda_{\mathrm{RL}}))(I).\end{aligned}$$

Since μ has concentration 1, the first term is at most $|\lambda_{\mathrm{LR}}| = \lambda_{\mathrm{L}}(t - x_\mu, x_\lambda]$. The measure μ_{R}, on the other hand, has mass 1, so, to prove the lemma, it suffices to show that the concentration of $\lambda_{\mathrm{LL}} + \lambda_{\mathrm{RL}}$ is at most $1 - \lambda_{\mathrm{L}}(t - x_\mu, x_\lambda]$.

By considering the various ways in which an interval of length 1 could overlap with $(t - x_\mu, x_\lambda]$, it is easy to see that

$$c(\lambda_{\mathrm{LL}} + \lambda_{\mathrm{LR}}) \le \max \{|\lambda_{\mathrm{LL}}|, |\lambda_{\mathrm{RL}}|, 1 - |\lambda_{\mathrm{LR}}|\}.$$

Now, the first and last of these terms are $|\lambda_{\mathrm{LL}}| = 1 - |\lambda_{\mathrm{LR}}| = 1 - \lambda_{\mathrm{L}}(t - x_\mu, x_\lambda]$, while

$$\begin{aligned}|\lambda_{\mathrm{RL}}| &= \lambda(t - x_\mu, t - x_\mu + 1) - \lambda_{\mathrm{L}}(t - x_\mu, x_\lambda] \\ &\le 1 - \lambda_{\mathrm{L}}(t - x_\mu, x_\lambda],\end{aligned}$$

since λ has concentration 1. Thus the result is proved. □

With this lemma, it is simple to deduce the following, more comprehensible version.

LEMMA 5. *If μ is a measure of mass $m \ge 1$ and λ is a measure of mass* 2, *and each has concentration at most* 1, *then the convolution $\mu * \lambda$ can be written as a sum of two measures of concentration at most* 1, *$\mu * \lambda = \nu' + \nu''$, with $|\nu'| = m + 1$ and $|\nu''| = m - 1$.*

Proof. With the same notation as in Lemma 4, set $\nu' = \nu$ and $\nu'' = \mu * \lambda - \nu$. Then, by that lemma, we have $\nu' \in \mathcal{M}(m + 1, 1)$, and certainly $\mu * \lambda = \nu' + \nu''$. Also, $\nu'' = \mu_{\mathrm{L}} * \lambda_{\mathrm{R}}$ and, since μ_{L} has concentration at most 1 while λ_{R} has mass 1, Lemma 3 shows that $\nu'' \in \mathcal{M}(m - 1, 1)$. □

These tools suffice for the proof of Theorem 1.

Proof of Theorem 1. Let μ_{X_i} be the distribution of X_i and set $\mu_i = 2\mu_{X_i}$. The theorem states that, whenever U is an open subset of $\mathbb{R}$ with diameter at most 1, then $(\otimes_1^n \mu_i)(U) \le \binom{n}{\lfloor n/2 \rfloor}$. More is true; in fact, $\otimes_1^n \mu_i$ can be written as a sum $\sum_1^s \nu_j$, of measures of concentration at most 1, where $(|\nu_j|)_1^s$ is a symmetric profile for $\mathscr{P}(n)$. Again, the proof goes by induction on n. If $\otimes_1^{n-1} \mu_i = \sum_1^s \nu_j$, where each ν_j has concentration 1 and $(|\nu_j|)_1^s$ is a symmetric profile of $\mathscr{P}(n-1)$, then Lemma 5 gives

$$\begin{aligned} \bigotimes_1^n \mu_i &= \left(\sum_1^s \nu_j\right) * \mu_n \\ &= \sum_1^s \nu_j * \mu_n \\ &= \sum_1^s \nu_j' + \nu_j''. \end{aligned}$$

Since each ν_j splits into two new measures, of masses $|\nu_j| + 1$ and $|\nu_j| - 1$, the masses of the new decomposition form a symmetric profile of $\mathscr{P}(n)$. So

$$\begin{aligned} c\left(\sum_1^n X_i\right) &= 2^{-n} c\left(\bigotimes_1^n \mu_i\right) \\ &= 2^{-n} c\left(\sum_{j=1}^s \nu_j\right) \\ &\leq 2^{-n} \sum_{j=1}^s c(\nu_j) \\ &\leq 2^{-n} s \\ &= 2^{-n} \binom{n}{\lfloor n/2 \rfloor}. \end{aligned}$$

Thus the result is proved. □

2. Concentration at most $1/q$. In this section, we extend Theorem 1 to measures of concentration at most $1/q$ for some fixed integer q. The techniques used are a straightforward extension of those in the proof of Theorem 1.

The first step is to note that the poset $[q]^n$ has a symmetric chain decomposition. This poset is ranked by *weight*: $w(x) = \sum_1^n x_i$ for $x \in [q]^n$. A chain $x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(r)}$ is *symmetric* if $w(x^{(k+1)}) = w(x^{(k)}) + 1$ and $w(x^{(1)}) + w(x^{(r)}) = n(q-1)$. It was proved by de Bruijn, Kruyswijk, and Tengbergen [1] that $[q]^n$ has a symmetric chain decomposition. Again, we say that a sequence $(m_j)_1^s$ is a *symmetric profile* for $[q]^n$ if, for some (and hence, up to rearrangement, for any) symmetric chain decomposition $(\mathscr{S}_j)_1^s$, we have $|\mathscr{S}_j| = m_j$. The required information about how symmetric profiles change is here stated as a lemma.

LEMMA 6. *Let $(m_j)_1^s$ be a symmetric profile for $[q]^{n-1}$, and, for each $j = 1, 2, \ldots, s$, set $r_j = \min \{q, m_j\}$. Then the sequence obtained by replacing m_j by the r_j values $m_j + q + 1 - 2k$ for $k = 1, 2, \ldots, r_j$ is a symmetric profile for $[q]^n$.*

Proof. Consider a chain $\mathscr{S} = (x^{(k)})_1^r$ belonging to a symmetric chain decomposition of $[q]^{n-1}$. For $x \in [q]^{n-1}$ and $h \in [q]$, denote by $x + he_n$ the element of $[q]^n$ formed by appending h to x. For $0 \leq l \leq \min(q, r) - 1$, let $\mathscr{S}^{(l)} = \{x^{(k)} + he_n : \min(r-k, h) = l\}$. Then each $\mathscr{S}^{(l)}$ is a symmetric chain in $[q]^n$, and the union of all the chains arising in this way forms a symmetric chain decomposition of $[q]^n$. See Fig. 1 and [1] for more details. □

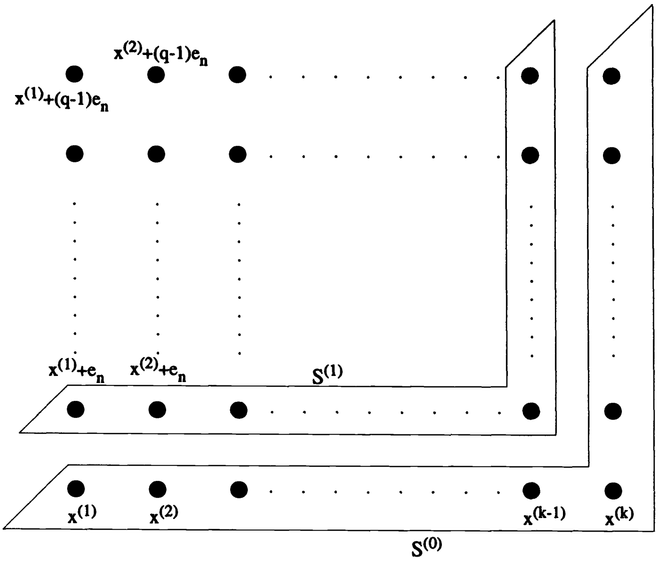


FIG. 1. *The construction of symmetric chains.*

The next lemma extends Lemma 5 to cover the present case. It states, in essence, that we can treat a measure of mass m and concentration 1 much like the poset $[m]$. In particular, the convolution of an element of $\mathcal{M}(m, 1)$ and one of $\mathcal{M}(l, 1)$ behaves similarly to the poset $[m] \times [l]$ and can be "peeled apart" in the same fashion. This peeling process will allow us to write a convolution $\otimes_1^n \mu_i$ (where $\mu_i \in \mathcal{M}(q, 1)$) as a sum of measures whose mass profile mimics that of a symmetric chain decomposition of the poset $[q]^n$.

LEMMA 7. *Given $m, l \in \mathbb{N}$, set $r = \min\{m, l\}$. If $\mu \in \mathcal{M}(m, 1)$ and $\lambda \in \mathcal{M}(l, 1)$, then $\mu * \lambda$ can be written as a sum $\sum_{k=1}^{r} \nu^{(k)}$ in which $\nu^{(k)} \in \mathcal{M}(m + l + 1 - 2k, 1)$.*

Proof. The case $l = 2$ of this lemma is precisely Lemma 5, and the case where $l = 1$ is trivial: the splitting $\nu^{(1)} = \mu * \lambda$ will do. The general case is proved by induction on l. Using the result and notation of Lemma 4, set $\nu^{(1)} = \mu_{\mathrm{R}} * \lambda + \mu * \lambda_{\mathrm{L}} - \mu_{\mathrm{R}} * \lambda_{\mathrm{L}}$, with $|\nu^{(1)}| = m + l - 1$ and $c(\nu^{(1)}) \leq 1$. After some judicious relabelling, the induction hypothesis states that $\mu * \lambda - \nu^{(1)} = \mu_{\mathrm{L}} * \lambda_{\mathrm{R}}$ can be written as $\sum_{k=2}^{r} \nu^{(k)}$, with $\nu^{(k)} \in \mathcal{M}(m + l + 1 - 2k, 1)$. □

Proof of Theorem 2. We prove by induction that the convolution $\otimes_1^n q\mu_{X_i}$ can be decomposed as a sum of measures of concentration at most 1 whose mass profile is a symmetric profile for $[q]^n$. Indeed, this is trivially possible if $n = 1$. For the induction step, let $(\nu_j)_1^s$ be a decomposition for $\otimes_1^{n-1} q_i\mu_{X_i}$. Then

$$\mu = \left(\sum_{j=1}^{s} \nu_j\right) * (q\mu_{X_n})$$
$$= \sum_{j=1}^{n} \nu_j * (q\mu_{X_n}).$$

Now, by Lemma 7, each convolution $\nu_j * (q\mu_{X_n})$ can be written as a sum $\sum_{k=1}^{r_j} \nu_j^{(k)}$, where $r_j = \min\{q, |\nu_j|\}$ and $\nu_j^{(k)} \in \mathcal{M}(|\nu_j| + q + 1 - 2k, 1)$ for $k = 1, 2, \ldots, r_j$. The collection of all nonzero $\nu_j^{(k)}$ is a decomposition of μ into measures of concentration 1, and by Lemma 6 their masses form a symmetric profile for $[q]^n$. □

Remark. The proof of Theorem 2 can easily be extended to prove rather more. Indeed, if $(X_i)_1^n$ are independent real-valued random variables with $c(X_i) \leq 1/q_i$ for all i (where $q_1, \dots, q_n$ are integers), we can show that the concentration of $\sum X_i$ is at most the proportion of $\prod_1^n [q_i]$ occupied by the largest layer. In fact, by using slightly more information about the symmetric profile of $\prod_1^n [q_i]$, we can show that, given r open intervals $(I_k)_1^r$, each of length at most 1, the probability that $\sum_1^n X_i$ lies in $\bigcup_1^r I_k$ is bounded by the proportion of $\prod_1^n [q_i]$ occupied by the r largest layers.

3. Other values of the concentration. What can be said about the sum of independent random variables of concentration at most c for values of c not of the form $1/q$? Might it be that the sum of n independent random variables, each of concentration p/q (p, q coprime integers) has concentration at most the proportion of $[q]^n$ occupied by the p largest layers? It is easy to see that this cannot be the case, since we may have quite complicated fractions p/q that closely approximate some simple number such as $\frac{1}{2}$. For instance, let X_1, X_2 be independent and identically distributed random variables with distribution $(\delta_0 + \delta_1)/2$. The X_i certainly have concentration at most $\frac{4}{7}$. However, their sum has concentration $\frac{1}{2}$, which is greater than $\frac{24}{49}$, the proportion of $[7]^2$ occupied by the four largest layers.

On the other hand, somewhat surprisingly given this simple example, the question above does have a positive answer when $p = 2$, in other words, for concentrations of the form $2/q$ with q odd. The proof proceeds by showing how to "peel apart" convolutions of measures of concentration 2.

One preliminary lemma is necessary.

LEMMA 8. *If $q \in \mathbb{N}$ and $\mu \in \mathcal{M}(q, 2)$, then there exist measures μ_0, μ_1, both of concentration* 1, *with $\mu = \mu_0 + \mu_1$ and $|\mu_0| = \lfloor q/2 \rfloor$ and $|\mu_1| = \lceil q/2 \rceil$.*

Proof. Split μ into parts of mass 1 and (almost) disjoint support going from left to right. In other words, define $\nu_1 = \text{Left}\,(\mu, 1)$ and for $j = 2, 3, \dots, q$ set

$$\nu_j = \text{Left}\left(\mu - \sum_1^{j-1} \nu_k, 1\right).$$

Now collect all the ν_j for j even together and similarly for all the odd ν_j as follows:

$$\mu_h = \sum_{j \equiv h (\text{mod } 2)} \nu_j, \qquad h = 0, 1.$$

Then it is clear that μ_0, μ_1 have the correct masses. Now let $I = (t, t + 1)$ be an arbitrary interval. Since μ has concentration 2, at most three of the ν_k can have $\nu_k(I) > 0$. This is because, if four of them could detect I, of necessity four consecutive ones, ν_k, ν_{k+1}, ν_{k+2}, and ν_{k+3}, say, then we would have $\nu_{k+1} = \nu_{k+2} = 1$ and $\mu(I) \geq (\sum_{l=0}^3 \nu_{k+l})(I) > (\nu_{k+1} + \nu_{k+2})(I) = 2$. This contradicts the fact that μ has concentration 2.

If exactly three of the ν_k give positive measure to I, then similarly they are ν_k, ν_{k+1} and ν_{k+2} with $\nu_{k+1}(I) = 1$. Thus

$$\nu_k(I) + \nu_{k+2}(I) = \mu(I) - \nu_{k+1}(I) \leq 2 - 1 = 1.$$

So, in the case when the support of three ν_k intersect I, we have $\mu_0(I)$, $\mu_1(I) \leq 1$. In the case when at most two supports are involved, it is clear that both μ_0 and μ_1 are at most 1 on I, and the proof is complete. □

We are very fortunate to have the following lemma.

LEMMA 9. *If $\mu \in \mathcal{M}(m, 2)$ and $\lambda \in \mathcal{M}(l, 2)$, with m and l odd, then $\mu * \lambda$ can be written as a sum of measures of concentration at most* 2 *whose masses form a symmetric profile for $[m] \times [l]$.*

Proof. Write $m = 2a + 1$ and $l = 2b + 1$. By Lemma 8, we can write $\mu = \mu_0 + \mu_1$ and $\lambda = \lambda_0 + \lambda_1$, measures of concentration at most 1 and masses a, $a + 1$ and b, $b + 1$, respectively. Without loss of generality, $b \leq a$, but there are two cases, when $b < a$ and when $b = a$.

Case 1. $b < a$.

By Lemma 7, $\mu_0 * \lambda_0$ splits into b measures of concentration at most 1 and masses $a + b - 1, a + b - 3, \ldots, a - b + 1$. The convolution $\mu_1 * \lambda_0$ can be decomposed into b measures of concentration at most 1 of masses $a + b, a + b - 2, \ldots, a - b + 2$. Summing in pairs, we can write $\mu * \lambda_0$ as the sum of b measures of concentration at most 2 of masses $2a + 2b - 1, 2a + 2b - 5, \ldots, 2a - 2b + 3$.

In a similar fashion, both $\mu_0 * \lambda_1$ and $\mu_1 * \lambda_1$ split into $b + 1$ pieces of concentration at most 1. When paired up, these give measure of concentration at most 2 and masses $2a + 2b + 1, 2a + 2b - 1, \ldots, 2a - 2b - 3$. The two collections together provide an appropriate splitting for $\mu * \lambda$.

Case 2. $b = a$.

Partition $\mu * \lambda_0$ as before. Now, however, $\mu_0 * \lambda_1$ splits into only a parts, of masses $a + b, a + b - 2, \ldots, 2$, whereas $\mu_1 * \lambda_1$ splits as before into $b + 1$ parts: masses $a + b + 1, a + b - 1, \ldots, 1$. Pair these measures, leaving the final measures of mass 1 unpaired. This produces b measures of concentration at most 2, of masses $2a + 2b + 1$, $2a + 2b - 3, \ldots, 5$. Together with the remaining measure of mass 1 (and therefore certainly of concentration at most 2), this gives us exactly the desired splitting. □

We are ready to study the case of concentration $2/q$.

THEOREM 10. *Let $q \in \mathbb{N}$ be odd and let $(X_i)_1^n$ be independent real-valued random variables with $c(X_i) \leq 2/q$. Let W_1 and W_2 be the sizes of the two largest layers in $[q]^n$. Then*

$$c\left(\sum_{i=1}^{n} X_i\right) \leq (W_1 + W_2)/q^n.$$

Proof. Following the proof of Theorem 1, using Lemma 9 rather than Lemma 5, we can write $\mu = \otimes_1^n (q_i \mu_{X_i})$ as a sum $\sum_1^s \nu_j$, where each ν_j has concentration at most 2 and $(|\nu_j|)_1^s$ is a symmetric profile for $[q]^n$. What does this profile look like? If $W_2 < W_1$, then there must be exactly $W_1 - W_2$ 1's in it. If $W_1 = W_2$, then the corresponding chain decomposition can have no chains of length 1. In summary, exactly $W_1 - W_2$ of the ν_j have mass 1 (and therefore concentration at most 1) and the other W_2 have concentration at most 2. Thus

$$\begin{aligned} c(\mu) &\leq \sum_1^s c(\nu_j) \\ &\leq 2W_2 + (W_1 - W_2) \\ &= W_1 + W_2, \end{aligned}$$

and so $c(X) = c(\mu)/q^n \leq (W_1 + W_2)/q^n$. □

We note that Theorem 10 is best possible, as may be seen by taking each X_i to have distribution $(1/q)(\delta_0 + \delta_1 + \cdots + \delta_{q-1})$.

The question remains as to what can be said for other values of the concentration. If $(X_i)_1^n$ are independent real-valued random variables each of concentration at most c, can we give good upper bounds for the concentration of $\sum X_i$?

4. The vector-valued case. In the first three sections, we have concentrated on the behavior of real-valued random variables. We turn now to the situation that was Jones's

[3] primary concern, the vector-valued case. Refer to [7] for general background and notation about normed spaces.

Jones studied the following question. A subset M of a normed space E is said to be 1-*separated* if, for any distinct $x, y \in M$, we have $\|x - y\| \geq 1$. Given sets $M_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,q}\}$ for $i = 1, 2, \ldots, n$ such that each is 1-separated, form the corresponding random variables $(X_i)_1^n$ with X_i having distribution $(1/q) \sum_{j=1}^{q} \delta_{x_{i,j}}$. Is it true that the concentration of $X = \sum X_i$ is always bounded as we would wish, by the proportion of $[q]^n$ occupied by the largest layer? This question remains unanswered.

In his study of the problem, Jones introduced some useful definitions (given here in slightly less generality than in [3]). Let M_1 and M_2 be 1-separated finite subsets of a normed space E with $|M_i| = m_i$. We say that the pair M_1, M_2 has the *B.T.K. chain property* if their sum $M_1 + M_2$, counted with multiplicities, can be partitioned into a family of 1-separated subsets whose size profile is a symmetric profile for $[m_1] \times [m_2]$. We say that E has the *B.T.K. chain property* if every pair of finite 1-separated subsets of E has the B.T.K. chain property. If E has the B.T.K. chain property, then the above question has an affirmative answer (as may be seen by mimicking symmetric chain decompositions of $[q]^n$).

A partial ordering $\leq$ on a normed space E is said to be *compatible* if it is translation invariant (i.e., satisfies $x \leq y$ if and only if $x + a \leq y + a$ for all $x, y, a \in E$) and has the property that distinct $x, y \in E$ are comparable if and only if $\|x - y\| \geq 1$. Thus, for example, $\mathbb{R}$ certainly has a compatible ordering: we let x precede y if $x + 1 \leq y$ in the usual order.

Jones showed that, if X has a compatible order, then X has the B.T.K. chain property. He proved that two-dimensional Hilbert space, and hence each higher-dimensional Hilbert space, fails to have the B.T.K. chain property and (a fortiori) has no compatible order. He asked whether l_∞ has a compatible order, or at least satisfies the B.T.K. chain property.

In some sense, Jones answered this question himself, since we can find two-dimensional subspaces of l_∞ isometric to Hilbert space, and so l_∞ cannot have the B.T.K. chain property. In fact, compatible orders are rather hard to find: no normed space of dimension greater than 1 has a compatible ordering. Moreover, the condition that $\leq$ be translation invariant is not the reason.

PROPOSITION 11. *Let E be a normed space of dimension greater than* 1. *Then there is no partial ordering $\leq$ on E such that distinct $x, y \in E$ are comparable if and only if* $\|x - y\| \geq 1$.

Proof. It clearly suffices to show that no two-dimensional example exists, so let us suppose that E is a two-dimensional space with such an ordering $\leq$. Let $\{x_1, x_2\}$, $\{x_1^*, x_2^*\}$ be an Auerbach system for X. Thus x_1 and x_2 have norm 1; x_1^* and x_2^*, belonging to E^*, have dual norm 1; and $x_i^*(x_j) = \delta_{ij}$ (such a system can easily be found—see, e.g., [7]). Consider first the set $M = \{0, x_1, x_2\}$. Since $\|x_1 - x_2\| \geq x_1^*(x_1 - x_2) = 1$, the set M is 1-separated and hence totally ordered by $\leq$.

CLAIM. *Either $x_1 \leq 0 \leq x_2$ or $x_2 \leq 0 \leq x_1$.*

Otherwise, we may suppose, without loss of generality, that $0 \leq x_1 \leq x_2$. Consider then $y = x_2/2$. We have that $\|x_1 - y\| \geq x_1^*(x_1 - y) = 1$, so either $x_1 > y$ or $x_1 < y$. In the first case, we have $x_2 > x_1 > y$, despite the fact that $\|x_2 - y\| = \frac{1}{2}$. In the second case, $y > x_1 > 0$, which again contradicts the condition on $\leq$ since $\|y\| = \frac{1}{2}$. So the claim is proved.

Exactly the same reasoning, applied to $M' = \{0, x_1, x_1 + x_2\}$, shows that x_1 must be $\leq$-between 0 and $x_1 + x_2$. Similarly, we must have $x_1 + x_2$ between x_1 and x_2 and also x_2 between 0 and $x_1 + x_2$. However, these four conditions are incompatible—the $\leq$-maximum of the four vectors $\{0, x_1, x_2, x_1 + x_2\}$ does not lie between two others. This contradiction establishes the nonexistence of $(E, \leq)$. □

Jones's main positive result was that, if M_1 and M_2 are both 1-separated subsets of Hilbert space, with M_2 having size at most 3, then the pair M_1, M_2 has the B.T.K. chain property. (We should note that Kleitman based his proof of Theorem A exactly on the fact that in any normed space, any pair M_1, M_2 of 1-separated subsets has the B.T.K. chain property if M_2 has size at most 2.)

This pleasant fact about Hilbert space does not, unfortunately, generalise to arbitrary normed spaces. We present here an example of a normed space E and two 1-separated subsets, each of size 3, not having the B.T.K. chain property. In fact, we find 1-separated sets M_1, M_2 of size 3 such that the sum $M_1 + M_2$, far from having a partition into 1-separated subsets of size 5, 3, and 1, does not even contain a 1-separated subset of size 5.

PROPOSITION 12. *There exists a normed space E and* 1-*separated subsets M_1, $M_2 \subset E$ such that $|M_1| = |M_2| = 3$ but $M_1 + M_2$ contains no* 1-*separated subset of size* 5.

Proof. We define a norm on $\mathbb{R}^4$ in such a way that the sets $M_1 = \{0, e_1, e_2\}$ and $M_2 = \{0, e_3, e_4\}$ satisfy the conclusion of the proposition. More exactly, we ensure that each distance marked on Fig. 2 is strictly less than 1, while both M_1 and M_2 are 1-separated. To ensure that the requisite vectors are short, we define our norm $\|\cdot\|$ by taking for its unit ball the absolute convex hull of these vectors. In other words, we take as the unit ball the set

$$B_{\|\cdot\|} = \text{abs-co}\,\{e_1 + e_3, (e_2 - e_1) + e_4, e_2 - e_4, e_2 - e_3,$$
$$e_1 + (e_4 + e_3), e_2 + (e_3 - e_4), (e_1 - e_2) + (e_4 - e_3)\}.$$

By definition, all these vectors have norm at most 1. Now we show that the vectors e_1, e_2, e_3, $e_1 - e_2$, $e_3 - e_4$ have norm strictly greater than 1 by exhibiting functionals of (dual) norm at most 1 taking large values at those vectors. For instance, for $\varepsilon > 0$ sufficiently small, the functional $f = (1 + \varepsilon, 0, -\varepsilon, -2\varepsilon)$ has dual norm at most 1, because it takes values at most 1 in absolute value at the extreme points of the $\|\cdot\|$ unit ball. However, $f(e_1) = 1 + \varepsilon > 1$, and therefore $\|e_1\| > 1$. In similar fashion, we can exhibit

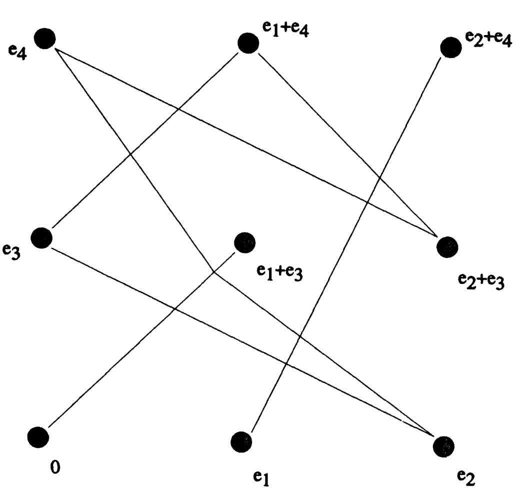


FIG. 2. *The pattern of small distances in Proposition* 12.

functionals to show that all the vectors we desire to be long are indeed long. For some small $\varepsilon > 0$, the following suffice:

$$\begin{aligned} e_1&: \quad (1+\varepsilon, 0, -\varepsilon, -2\varepsilon),\\ e_2&: \quad (3\varepsilon, 1+\varepsilon, \varepsilon, 2\varepsilon),\\ e_3&: \quad (-\varepsilon, \varepsilon, 1+\varepsilon, 3\varepsilon),\\ e_4&: \quad (2\varepsilon, \varepsilon, 3\varepsilon, 1+\varepsilon),\\ e_2-e_1&: \quad \tfrac{1}{2}(1+\varepsilon, -1, 2\varepsilon, \varepsilon),\\ e_3-e_4&: \quad \tfrac{1}{2}(-\varepsilon, \varepsilon, -1-\varepsilon, 1). \end{aligned}$$

The norm $\|\cdot\|$ does not behave exactly as we would like—the norms from Fig. 2 are at most 1, rather than strictly less than 1—but for some $0 < \lambda < 1$, the norm $\lambda\|\cdot\|$ will do. It is easy to check, from Fig. 2, that $M_1 + M_2$ contains no 1-separated subset of size 5. □

There are still many unanswered questions concerning the vector-valued case. The most striking and interesting one, it seems to us, is whether the following conjecture is true.

CONJECTURE 13. *Let E be a normed space and let $(X_i)_1^n$ be independent E-valued random variables of concentrations at most $\frac{1}{2}$. Then the concentration of $\sum X_i$ is at most $\binom{n}{\lfloor n/2 \rfloor}2^{-n}$.*

REFERENCES

[1] N. G. DE BRUIJN, D. K. KRUYSWIJK, AND CA. VAN EBBENHORST TENGBERGEN, *On the set of divisors of a number*, Nieuw Arch. Wisk., 23 (1952), pp. 191–193.

[2] P. ERDÖS, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc., 51 (1945), pp. 898–902.

[3] L. JONES, *On the distribution of sums of vectors*, SIAM J. Appl. Math., 34 (1978), pp. 1–6.

[4] G. O. H. KATONA, *On a conjecture of Erdös and a stronger form of Sperner's theorem*, Studia Sci. Math. Hungar., 1 (1966), pp. 59–63.

[5] D. J. KLEITMAN, *On a lemma of Littlewood and Offord on the distribution of linear combinations of vectors*, Adv. in Math., 5 (1970), pp. 155–157.

[6] ———, *On a lemma of Littlewood and Offord on the distribution of certain sums*, Math. Z., 90 (1965), pp. 251–259.

[7] J. LINDENSTRAUSS AND L. TZAFRIRI, *Classical Banach Spaces* I, Springer-Verlag, Berlin, 1977.

[8] J. E. LITTLEWOOD AND C. OFFORD, *On the number of real roots of a rnndom algebraic equation* (III), Math. USSR-Sb., 12 (1943), pp. 277–285.

[9] E. SPERNER, *Ein Satz über Untermengen einer endlichen Menge*, Math. Z., 27 (1928), pp. 544–548.

SIAM J. DISC. MATH.
Vol. 7, No. 1, pp. 102-107, February 1994

010

THE COMBINATORICS OF PERFECT AUTHENTICATION SCHEMES*

CHRIS MITCHELL†, MICHAEL WALKER‡, AND PETER WILD§

Abstract. The purpose of this paper is to prove the equivalence of perfect authentication schemes and maximum distance separable codes.

Key words. authentication schemes, incidence structures, maximum distance separable codes

AMS subject classifications. 05B05, 94A60, 94B65

1. Introduction. In this paper, we consider the following communications scenario, which involves an originator of messages, a recipient of messages, and a third party called the spoofer. The originator wishes to send a sequence $s_1, \ldots, s_n$ of n distinct source messages to the recipient. To enable the recipient to verify the authenticity of these messages, the originator encodes them, prior to transmission, into a sequence of encoded messages $m_1, \ldots, m_n$, using one of a finite set of encoding rules that is agreed in advance with the recipient. The recipient verifies the authenticity of the message m_i by checking that it is a valid encoding of s_i for the agreed encoding rule e. The spoofer observes the sequence of encoded messages $m_1, \ldots, m_n$ and attempts to construct a correctly encoded message for a different source message. That is, he attempts to find a message m that is the encoding under the (to him unknown) encoding rule e of some source message that is distinct from $s_1, \ldots, s_n$.

In [6] one of the authors established an information-theoretic lower bound for the expected probability $P(n)$ that the spoofer succeeds in this task and for the arithmetic mean of $P(0), P(1), \ldots, P(N)$, where N is the maximum length of the sequence that the originator might be required to send using the same encoding rule. In addition, necessary and sufficient conditions on the encoding scheme are derived that ensure that these bounds are met, and these conditions lead to the concept of an N-perfect authentication scheme.

Associated with an authentication scheme is an incidence structure, and the conditions that ensure that the information-theoretic bounds are met are reflected in structural requirements on this incidence structure. In this paper, we characterise the incidence structures associated with N-perfect authentication schemes and thereby prove that perfect authentication schemes are equivalent to maximum distance separable codes.

The theorem proved in this paper extends a result in [3], where it is shown that an incidence structure is associated with a 1-perfect authentication scheme if and only if it is a net. It also establishes the converse of the observation made in [6], and independently by Stinson in [5], that an MDS code (or, equivalently, a transversal design) may be used to construct a perfect authentication scheme.

2. Authentication schemes, incidence structures, and codes. An *authentication scheme* is a triple $\underline{A} = \underline{A}(S, M, E)$ of finite sets S, M, and E, where each element of E is an injective function of S into M, and each element in M is the image under this set

* Received by the editors on February 26, 1990; accepted for publication (in revised form) November 13, 1992.

† Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England.

‡ Vodafone Limited, The Courtyard, 2-4 London Road, Newbury, Berkshire RG13 1JL, England.

§ Department of Mathematics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England.

of functions of precisely one element in S. The elements of S are known as *source messages,* those of M are called *encoded messages*, and the functions in E are called *encoding rules*.

An authentication scheme as defined above is often referred to in the literature as a nonsplitting, Cartesian scheme (see [4]). The nonsplitting property means that, once an encoding rule has been selected, then the encoded message for each source message is unambiguously defined. For each encoding rule $e \in E$ and each source message $s \in S$, we denote by $m = e(s)$ the encoded message for s produced by e. The Cartesian property means that there is no secrecy in the scheme, in the sense that, if an encoded message is observed, then there is no ambiguity about which source message it encodes, even if the encoding rule is unknown. This is a consequence of the requirement that each encoded message is the encoding of precisely one source message. We use the notation $s = S(m)$ to denote the unique source message corresponding to the encoded message m.

With an authentication scheme $\underline{A}$, we may associate an incidence structure $\underline{I}(\underline{A}) = I(E, M, I)$. The set of points of this incidence structure is E, the set of blocks is M, and the incidence relation I is defined by the rule

$$eIm \quad \text{if and only if } m = e(S(m)).$$

That is, point e is incident with block m precisely when the encoded message m is obtained by encoding $S(m)$ under the encoding rule e. We use standard notation for incidence structures, as may be found, for instance, in [1]. Thus we denote by $(m) = \{e \in E| e(S(m)) = m\}$ the set of all points incident with the block m, and by $(e) = \{m \in M| e(S(m)) = m\}$ the set of all blocks incident with the point e. Moreover, we extend this notation by defining $(s) = \{m \in M | S(m) = s\}$ to be the set of all encoded messages that are encodings of the source s. Finally, we denote by $[x]$ the cardinality of the set (x).

The incidence structure $\underline{I} = \underline{I}(\underline{A})$ enjoys the properties that $\{(s)|s \in S\}$ is a partition of M, and, for each $s \in S$, the set $\{(m)|m \in (s)\}$ is a partition of E. That is, $\{(s)|s \in S\}$ is a parallelism of $\underline{I}$, where we recall that a parallelism of an incidence structure is a partition of its blocks into classes with the property that each point of the structure is incident with precisely one block from each of the classes. The property of having a parallelism actually characterises those incidence structures that are associated with authentication schemes as described above.

To see this, let $\underline{I} = \underline{I}(P, B, I)$ be an incidence structure, with points P and blocks B, which possesses a parallelism S. To avoid complications, assume that $\underline{I}$ does not have repeated points; that is, if $(p) = (p')$, then $p = p'$. We use each point $p \in P$ to define a function from S into B as follows: For each $s \in S$, set $p(s)$ to be the unique block in the class s that is incident with p. Then it is trivial to check that $\underline{A} = \underline{A}(S, B, P)$ is an authentication scheme and that $\underline{I}(\underline{A}) = \underline{I}$.

Having defined and characterised the incidence structure associated with an authentication scheme, we now turn to considering codes for authentication schemes. The approach we take is via the associated incidence structure. Although there are a number of other ways of associating codes and authentication schemes, this is the most convenient for our purposes.

We begin by recalling that a code $\underline{C}$ of length r over a finite alphabet A is a nonempty set of r-tuples with entries in A. The elements $c = (c_1, \dots, c_r)$ of $\underline{C}$ are called codewords.

With any code $\underline{C}$, we can associate an incidence structure $\underline{I} = \underline{I}(\underline{C})$, which has a parallelism. The points of $\underline{I}$ are the codewords c. The blocks of $\underline{I}$ are the pairs (i, a), where $i \in \{1, \dots, r\}$, $a \in A$, and a is the ith entry of at least one codeword. Incidence is then defined by the rule

$$cI(i, a) \quad \text{if and only if } c_i = a.$$

Now if, for each $i \in \{1, \ldots, r\}$ we define (i) to be the set of all blocks of the form (i, a), then $\{(i) \mid i \in \{1, \ldots, r\}\}$ is a parallelism of $\underline{I}$.

Conversely, to any incidence structure $\underline{I}$ with a parallelism, we can associate a code $\underline{C}$ such that $\underline{I}(\underline{C}) = \underline{I}$. To see this, let $\underline{I} = \underline{I}(P, B, I)$ be an incidence structure with parallelism S. Let $r = |S|$ and label the parallel classes $1, \ldots, r$. For each parallel class $i \in \{1, \ldots, r\}$, let φ_i be an injection from (i) into a suitably large set A and identify block $b \in (i)$ with the pair $(i, \varphi_i(b))$. For each point $p \in P$, define codeword $(p_1, \ldots, p_r)$ by $p_i = \varphi_i(b)$, where b is the unique block in the parallel class i incident with p. Then the set $\underline{C} = \{(p_i, \ldots, p_r) \mid p \in P\}$ is a code of length r over A and $\underline{I} = \underline{I}(\underline{C})$.

Combining the observations of this section, we see that, to any authentication scheme $\underline{A} = \underline{A}(S, M, E)$, we can associate a code $\underline{C} = \underline{C}(\underline{A})$. This code has length $|S|$, contains $|E|$ codewords, and is defined over an alphabet A of size equal to the maximum of the number of encodings of a source message. Conversely, given a code $\underline{C}$, we can construct an authentication scheme $\underline{A}$ with $\underline{C}(\underline{A}) = \underline{C}$.

3. MDS codes and perfect authentication schemes. Let $\underline{C}$ be a code of length r over a finite alphabet A of cardinality q. Then $\underline{C}$ is a *maximum distance separable* (MDS) code if and only if, for some t, it satisfies the following condition: Given any t distinct positions $i_1, \ldots, i_t$ and any sequence $a_1, \ldots, a_t$ of not necessarily distinct elements of A, there is exactly one codeword $c = (c_1, \ldots, c_r) \in \underline{C}$ with $c_{i_j} = a_j$ for $j = 1, \ldots, t$.

We refer to a code of this type as an MDS code with parameters (r, t, q). For further information on MDS codes, refer to [2]. It should be noted that our notation (r, t, q) is different from that used in [2].

We need the following characterisation of MDS codes in terms of their associated incidence structures as defined in the last section. The result is straightforward to prove using counting arguments and is therefore presented without proof. We use the following extension to the notation established for incidence structures in the last section. Let $\underline{b} = (b_1, \ldots, b_j)$ be a sequence of j blocks of an incidence structure. Then $(\underline{b})$ is the set of points that are incident with all the blocks $b_1, \ldots, b_j$, and $[\underline{b}]$ is the cardinality of this set.

LEMMA 3.1. *Let $\underline{I} = \underline{I}(\underline{C})$ be the incidence structure associated with an* MDS *code with parameters (r, t, q), let $0 \le j \le t$, and let $\underline{b} = (b_1, \ldots, b_j)$ be a sequence of j blocks belonging to different parallel classes. Then*

$$[\underline{b}] = q^{t-j}.$$

Conversely, if $\underline{I}$ is an incidence structure with a parallelism that satisfies this condition for some q, some t, and all $0 \le j \le t$, then $\underline{C} = \underline{C}(\underline{I})$ is MDS *with parameters (r, t, q), where r is the number of paralleled classes of $\underline{I}$.*

COROLLARY 3.2. *If an incidence structure satisfies the conditions of Lemma* 3.1, *then each parallel class contains exactly q blocks.*

We use this lemma to establish the equivalence of MDS codes and perfect authentication schemes. We begin by recalling the definition and characteristic properties of a perfect authentication as presented in [6]. To do this, we must first review the information theoretic measure of the security of an authentication scheme. For a more detailed discussion of the concepts, refer to [6].

Let $\underline{A} = \underline{A}(S, M, E)$ be an authentication scheme. To use this scheme, an originator and recipient of messages share an encoding rule e, which is selected from E according to some probability distribution $p(e)$. When the originator wishes to communicate a source message, he encodes it using e and sends the encoded message $m = e(s)$. Upon receiving m, the recipient validates it by using e to confirm that $e(S(m)) = m$. Suppose that the originator sends a sequence $m_1, \ldots, m_n$ of distinct encoded messages, all produced using the same encoding rule e, and that these messages are observed by a spoofer. The spoofer attempts to construct the correct encoding under e for some source message distinct from $S(m_1), \ldots, S(m_n)$. It is assumed that the spoofer has knowledge of $\underline{A}$ and plays the best strategy open to him. All he does not a priori know is the particular encoding rule e. We denote by $P(n)$ the expected probability that the spoofer succeeds in his task. We assume that n is not allowed to exceed some maximum value N and we let P_N be the arithmetic mean of $P(0), P(1), \ldots, P(N)$. The following lower bound for P_N is proved in [6]:

$$-\log P_N \leq H(E)/(N+1) \leq \log |E|/(N+1),$$

where $H(E)$ is the entropy of the distribution $p(e)$. These inequalities lead to the definition of an N-perfect authentication scheme.

An authentication scheme is said to be the *N-perfect* if $-\log P_N = H(E)/(N+1)$ with $p(e)$ the uniform distribution (so that, in this case, $H(E) = \log |E|$). Necessary and sufficient conditions for an authentication scheme to be N-perfect are given in [6]. These conditions form the basis for the main theorem of this paper and are summarised below in Lemma 3.3. To state these conditions and prove our theorem, it is first necessary to model the way in which source and encoded messages are generated and also to introduce more notation.

The sequence of source messages $s_1 = S(m_1), \ldots, s_n = S(m_n)$ generated by the originator and observed by the spoofer is modelled by a stochastic process $p(s_1, \ldots, s_n)$. We denote by $p(s_{n+1} | s_1, \ldots, s_n)$ the probability that the spoofer selects source message s_{n+1} with which to launch his attack, given that he has observed $s_1, \ldots, s_n$. All source messages are assumed to be distinct, so the process satisfies $p(s_j | s_1, \ldots, s_{j-1}) = 0$ whenever $s_j \in \{s_1, \ldots, s_{j-1}\}$. We also assume the converse. Thus our process satisfies

$$p(s_j | s_1, \ldots, s_{j-1}) = 0 \quad \text{if and only if } s_j \in \{s_1, \ldots, s_{j-1}\}.$$

We assume that the selection of the encoding rule is independent of the process that generates the source messages. Thus the probability that a sequence $\underline{m} = (m_1, \ldots, m_n)$ of encoded messages is produced by the originator and observed by the spoofer is given by

$$p(\underline{m}) = p(S(\underline{m}))p((\underline{m})),$$

where we use the notation

$$S(\underline{m}) = (S(m_1), \ldots, S(m_n)) \quad \text{and} \quad (\underline{m}) = \{e \in E | e(S(m_j)) = m_j, j = 1, \ldots, n\}.$$

The following additional notation will be used in the statement of Lemma 3.3 and in the proof of Theorem 3.4. Let $\underline{m} = (m_1, \ldots, m_n)$ be such that $p((\underline{m})) \neq 0$ and let $s \in S$. For each $m \in (s)$, define

$$(s | \underline{m}) = \{m \in (s) | p((m) | (\underline{m})) \neq 0\}.$$

Thus, in terms of the incidence structure $\underline{I}$ associated with $\underline{A}$, the set $(s | \underline{m})$ consists of all those blocks in the parallel class (s) that are potential valid encodings for s, given that the sequence $\underline{m}$ of encoded messages has already been produced. Observe that, if $n = 0$ so that $\underline{m}$ is the empty sequence, then $(s | \underline{m})$ consists of all those blocks of the incidence structure $\underline{I}(\underline{A})$ that belong to the parallel class (s) and are incident with at least one point $e \in E$ for which $p(e) \neq 0$.

LEMMA 3.3. *An authentication scheme $\underline{A} = \underline{A}(S, M, E)$ is N-perfect if and only if, for every $0 \leq n \leq N$, the following holds: If $\underline{m} = (m_1, \ldots, m_n)$ with $p(\underline{m}) \neq 0$, if $s \in S$ with $p(s|S(\underline{m})) \neq 0$, and if $m \in (s|\underline{m})$, then*

$$\log |E|/(N+1) = H(E)/(N+1) = -\log p((m)|(\underline{m})) = \log |(s|\underline{m})|.$$

The proof of the lemma follows immediately from [6, Thm. 2] and the definition of N-perfect. With the help of this result and Lemma 3.1, we may now prove our main theorem.

THEOREM 3.4. *Let $\underline{C} = \underline{C}(\underline{A})$ be the code associated with an N-perfect authentication scheme $\underline{A}$ with source messages S. Then $\underline{C}$ is an* MDS *code with parameters $(|S|, N+1, q)$, where $q = [s]$ for all $s \in S$. Conversely, if $\underline{C}$ is an* MDS *code with parameters $(r, N+1, q)$, then $\underline{C} = \underline{C}(\underline{A})$ for some N-perfect authentication scheme $\underline{A}$ with r source messages and q encodings for each source message.*

Proof. Let $\underline{A} = \underline{A}(S, M, E)$ be an N-perfect authentication scheme and let $\underline{I} = I(\underline{A})$ be the incidence structure associated with it. Applying Lemma 3.3 with $n = 0$ yields

$$\log [s] = \log |E|/(N+1) \quad \text{for all } s \in S.$$

Thus the number of blocks in each parallel class of $\underline{I}$ is a constant q, and $|E| = q^{N+1}$. We show that, if $0 \leq j \leq N+1$ and if $\underline{m} = (m_1, \ldots, m_j)$ is a sequence of blocks of $\underline{I}$ belonging to different parallel classes, then $[\underline{m}] = q^{N+1-j}$. The first part of the theorem then follows directly from Lemma 3.1.

To prove the above statement, we proceed by induction on j. We have already proved the result for $j = 0$, so we assume that $1 \leq j \leq N+1$, and the statement is true for $j - 1$. Let $\underline{m}' = (m_1, \ldots, m_{j-1})$ and consider the terms in the identity

$$p((\underline{m})) = p((m_j)|(\underline{m}'))p((\underline{m}')).$$

Since $p(e)$ is uniform, and, as $|E| = q^{N+1}$, we have $p((\underline{m})) = [\underline{m}]q^{-(N+1)}$ and similarly for $p((\underline{m}'))$. Thus $[\underline{m}] = p((m_j)|(\underline{m}'))[\underline{m}']$. Applying the induction hypothesis to $[m']$ now gives

$$(1) \qquad [\underline{m}] = p((m_j)|(\underline{m}'))q^{N+1-(j-1)}.$$

Now consider the term $p((m_j)|(\underline{m}'))$. First, we apply Lemma 3.3 with $n = j - 1$ to $\underline{m}'$ and $S(m_j)$. This gives

$$\log |(S(m_j)|\underline{m}')| = \log |E|/(N+1) = \log q.$$

However, we know that $[S(m_j)] = q$, so $(S(m_j)|\underline{m}') = (S(m_j))$. Thus $m_j \in (S(m_j)|\underline{m}')$, and Lemma 3.3 tells us that $-\log p((m_j)|(\underline{m}')) = \log q$. Thus $p((m_j)|(\underline{m}')) = q^{-1}$, and substitution of this in (1) proves $[\underline{m}] = q^{N+1-j}$, as required.

To prove the final part of the theorem, let $\underline{I} = \underline{I}(\underline{C})$ be the incidence structure associated with the MDS code and let $\underline{A} = \underline{A}(\underline{I})$ be the authentication scheme associated with $\underline{I}$. Let $p(e)$ be the uniform distribution on the set E of points of $\underline{I}$ and let $p(s_1, \ldots, s_n)$ be a process defined on the parallel classes S of $\underline{I}$ that satisfies $p(s_j|s_1, \ldots, s_{j-1}) = 0$ if and only if $s_j \in \{s_1, \ldots, s_{j-1}\}$. We prove that $\underline{A}$ is N-perfect by using Lemma 3.1 to show that the conditions of Lemma 3.3 are satisfied.

To this end, let $0 \leq n \leq N$, let $\underline{m} = (m_1, \ldots, m_n)$ be such that $p(\underline{m}) \neq 0$, let $s \in S$ with $p(s|S(\underline{m})) \neq 0$, and let $m \in (s|\underline{m})$. Since $p(\underline{m}) = p(S(\underline{m}))p(\underline{m})) \neq 0$, it follows that the parallel classes $S(m_1), \ldots, S(m_n)$ are distinct. Moreover, since

$p(s \mid S(\underline{m})) \neq 0$, the parallel class s is distinct from $S(m_1), \ldots, S(m_n)$. Thus, from Lemma 3.1, we have

$$[\underline{m}] = q^{N+1-n} \quad \text{and} \quad [\underline{m}, m] = q^{N+1-n-1}.$$

Thus

$$p((m) \mid (\underline{m})) = [\underline{m}, m]/[\underline{m}] = q^{-1}.$$

It follows that $(s \mid \underline{m}) = (s)$. However, $H(E) = \log |E|$, $|E| = q^{N+1}$ by Lemma 3.1, and $[s] = q$ by Corollary 3.2; so all the equalities in Lemma 3.3 hold.

REFERENCES

[1] P. Dembowski, *Finite Geometries*, Springer-Verlag, Berlin, Heidelberg, 1968.

[2] F. J. Mac Williams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1977.

[3] M. De Soete, K. Vedder, and M. Walker, *Cartesian authentication schemes*, in Advances in Cryptology—Proc. of Eurocrypt 89, Lecture Notes in Computer Science, Vol. 434, Springer-Verlag, Berlin, 1990, pp. 476–490.

[4] G. J. Simmons, *Authentication theory/coding theory*, in Advances in Cryptology—Proc. of Crypto 84, Lecture Notes in Computer Science, Vol. 196, Springer-Verlag, Berlin, 1985, pp. 411–431.

[5] D. R. Stinson, *The combinatorics of authentication and secrecy codes*, J. Cryptology, 2 (1990), pp. 23–49.

[6] M. Walker, *Information-theoretic bounds for authentication schemes*, J. Cryptology, 2 (1990), pp. 131–143.

SIAM J. Disc. Math.
Vol. 7, No. 1, pp. 108-118, February 1994

FAULT-TOLERANT CIRCUIT-SWITCHING NETWORKS*

NICHOLAS PIPPENGER† AND GENG LIN‡

Abstract. The authors consider fault-tolerant circuit-switching networks under a random switch failure model. Three circuit-switching networks of theoretical importance—nonblocking networks, rearrangeable networks, and superconcentrators—are studied. The authors prove lower bounds for the size (the number of switches) and depth (the largest number of switches on a communication path) of such fault-tolerant networks and explicitly construct such networks with optimal size $\Theta(n(\log n)^2)$ and depth $\Theta(\log n)$.

Key words. nonblocking networks, rearrangeable networks, superconcentrator

AMS subject classifications. 94C15, 68E10, 05C35

1. Introduction.

In this paper, we study some fault-tolerant circuit-switching networks under a random switch failure model. In this model, each electrical switch in the network is independently in one of the following three states: (1) *open failure* (the switch is permanently *off* and fails to be *on*) with probability $0 < \varepsilon_1 < \frac{1}{2}$, (2) *closed failure* (the switch is permanently *on* and fails to be *off*) with probability $0 < \varepsilon_2 < \frac{1}{2}$, and (3) *normal state* (the switch functions correctly) with probability $1 = \varepsilon_1 - \varepsilon_2$. For simplicity of notation, we assume that $\varepsilon_1 - \varepsilon_2 = \varepsilon$. The measure of fault tolerance is the probability of the network fulfilling the communication task in the presence of switch failures. This model is essentially equivalent to that of Moore and Shannon [MS], who introduced it in the context of relay circuits computing Boolean functions. The model retains its relevance, since open and closed failures represent the two dominant failures modes both for metallic-contact switches (still frequently used, especially for video switching) and MOSFETs (metal-oxide semiconductor field-effect transistors), a common switching element in VLSI circuits.

2. The networks.

The circuit-switching networks we study in this paper are *nonblocking networks*, *rearrangeable networks*, and *superconcentrators*. Nonblocking networks were introduced by Clos [Cl] in 1953 to epitomize the activity of telephone communication. Beneš [B] in 1964 described the rearrangeable network. Rearrangeable networks are useful architectures for parallel machines. Aho, Hopcroft, and Ullman [AHU] in 1974 posed the problem of *superconcentrators*. Although their purpose was to hope to use them to establish a nonlinear lower bound for the Boolean circuit complexity of multiplication, superconcentrators proved to be central in a number of communication networks. For example, superconcentrators provide support for the task queue scheme (see [Co]) in parallel computing. Tremendous efforts on these networks have been made, and significant results obtained.

In this paper, we describe a circuit-switching network in terms of an acyclic directed graph. *Terminals* of the network (wires that connect the network to the outside world) are represented by distinguished vertices called *inputs* and *outputs*. Electrical links are represented by vertices other than inputs and outputs, and switches (single-pole single throw, connecting two links) by edges between the two corresponding vertices. The three states of a switch in the random switch failure model are therefore interpreted as (1) the edge ceases to exist (open failure), (2) two vertices of the edge contract to one (closed

* Received by the editors April 20, 1992; accepted for publication (in revised form) October 15, 1992.

† Department of Computer Science, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada, (nicholas@cs.ubc.ca).

‡ Bell-Northern Research, Box 3511, Station C, Ottawa, Ontario K1Y 4H7, Canada, (geng@bnr.ca).

failure), and (3) the edge is unaffected (normal state). In this paper, we say "graph" and "network" without distinction, and the same is true for "edge" and "switch."

Given a directed graph with n inputs and n outputs, it is said to be a "*nonblocking n-network*" if, given any set of vertex-disjoint paths from inputs to outputs and given any input and output not involved in these established paths, a new path that is vertex-disjoint from the established paths can be found from the requesting input to the requesting output; it is said to be a "*rearrangeable n-network*" if, given any one-to-one correspondence between the inputs and the outputs, there exists a set of n vertex-disjoint paths joining each input to its corresponding output; it is said to be an "*n-superconcentrator*" if, for every $r \leq n$, every set of r inputs, and every set of r outputs, there exists a set of r vertex-disjoint paths from the given inputs to the given outputs. It is obvious that a nonblocking n-network is a rearrangeable n-network, and a rearrangeable n-network is an n-superconcentrator.

The networks considered in this paper are based on directed graphs and distinguish the roles of inputs and outputs as terminals. Variants of these definitions exist for networks based on undirected graphs, and for which there is but one class of terminals. Our definition of "nonblocking" is also referred to as "strictly nonblocking," to distinguish it from the somewhat weaker notion of "wide-sense nonblocking" that also appears in the literature. The networks we call "rearrangeable" are sometimes referred to as "permutation" networks, though the latter term is also used for some variants of this notion.

The measures of complexity applied to such networks are *size* (the number of edges) and *depth* (the largest number of edges on any path from an input to an output). Shannon [S] showed an $\Omega(n \log n)$ size lower bound of rearrangeable n-networks. Beneš [B] presented an $O(n \log n)$ size and $O(\log n)$ depth construction for rearrangeable n-networks. The existence of $O(n \log n)$ size and $O(\log n)$ depth nonblocking n-networks was proved by Bassalygo and Pinsker [BP]. For n-superconcentrators, an $\Omega(n)$ size lower bound is obvious, and Valiant [V] showed an $O(n)$ size upper bound.

3. Fault tolerance. Given $0 < \varepsilon < \frac{1}{2}$, consider a network N subject to the random switch failure model. Let the event space Ω be the set of all graphs obtained from N. The probability measure on each graph is assigned in accordance to the number of failed edges. More precisely, if a graph $G \in \Omega$ has k failed edges, the probability that the random instance of N equals G is $(2\varepsilon)^k(1 - 2\varepsilon)^{n-k}$, where n is the number of edges in N. Given $0 < \delta < 1$, we say that N is an (ε, δ)-*nonblocking n-network* if the probability that the random instance of N contains a nonblocking n-network consisting of edges of normal state is greater than $1 - \delta$. Similarly, we define an (ε, δ)-*n-rearrangeable network* and an (ε, δ)-*n-superconcentrator*. We observe that an (ε, δ)-nonblocking n-network is an (ε, δ)-rearrangeable n-network, and an (ε, δ)-rearrangeable n-network is an (ε, δ)-n-superconcentrator. It is clear that, by choosing arbitrarily small δ, an (ε, δ)-nonblocking n-network or an (ε, δ)-rearrangeable n-network or an (ε, δ)-n-superconcentrator can fulfill its communication task with arbitrarily high probability.

The goal of this paper is to analyze the asymptotic behaviors of the size and depth of the (ε, δ)-nonblocking n-network, the (ε, δ)-rearrangeable n-network, and the (ε, δ)-*n-superconcentrator*. For this purpose, the exact values of $0 < \varepsilon < \frac{1}{2}$ and $0 < \delta < 1$ do not matter. To see this, we first need a result of Moore and Shannon [MS].

Define an $(\varepsilon, \varepsilon')$-1-*network* to be a directed graph with two distinguished vertices called *input* and *output*, in which each edge is randomly and independently subject to closed and open failures with probabilities of ε, respectively, and in which the probabilities that the input and the output contract into one vertex and that there is no path from the input to the output are both less than ε'.

PROPOSITION 1 (Moore and Shannon). *Given $0 < \varepsilon < \frac{1}{2}$ and $0 < \varepsilon' \leq \varepsilon$, there is an explicit construction of an $(\varepsilon, \varepsilon')$-1-network with $c_\varepsilon(\log_2(1/\varepsilon'))^2$ edges and $d_\varepsilon \log_2(1/\varepsilon')$ depth, where c_ε and d_ε are constants depending only on ε.*

To observe the fact that the exact value of ε does not affect the asymptotic behaviors of the size and depth, we suppose that $0 < \varepsilon_1 \leq \varepsilon_2 < \frac{1}{2}$ and that Φ is an (ε_1, δ)-n-superconcentrator with size L and depth D, for some $\delta < 1$. By Proposition 1, there is an $(\varepsilon_2, \varepsilon_1)$-1-network Ψ of size a and depth b (a and b are depending only on ε_2). The result of substituting this network Ψ for each edge in Φ is clearly an (ε_2, δ)-n-superconcentrator with size at most aL and depth at most bD. Similar arguments apply to (ε, δ)-rearrangeable n-networks and (ε, δ)-nonblocking n-networks as well.

To see the invariance with respect to the value of δ, we suppose that $0 < \delta_1 \leq \delta_2 < 1$ and that Φ is an (ε, δ_2)-n-superconcentrator, for some $\varepsilon < \frac{1}{2}$. The failure probability of Φ is a polynomial in ε and the constant term of this polynomial vanishes (since the network does not fail unless some switch fails). If we replace ε by $\varepsilon\delta_1/\delta_2$, every term in this polynomial decreases to at most δ_1/δ_2 times its previous value. Thus Φ is also an $(\varepsilon\delta_1/\delta_2, \delta_1)$-$n$-superconcentrator. Again, substitute each edge in Φ by an $(\varepsilon, \varepsilon\delta_1/\delta_2)$-1-network and the resulting network is an (ε, δ_1)-n-superconcentrator with the size and depth being affected by only a constant factor. Similar arguments apply to (ε, δ)-rearrangeable n-networks and (ε, δ)-nonblocking n-networks as well.

4. Main result and the overall strategy. In this paper, we show that the size and depth of (ε, δ)-n-superconcentrators, (ε, δ)-rearrangeable n-networks, and (ε, δ)-nonblocking n-networks are $\Theta(n(\log n)^2)$ and $\Theta(\log n)$.

The overall strategy is that we prove the $\Omega(n(\log n)^2)$ and $\Omega(\log n)$ lower bounds for size and depth of a $(\frac{1}{4}, \frac{1}{2})$-$n$-superconcentrator, and we construct $(10^{-6}, \delta)$-nonblocking n-networks with $O(n(\log n)^2)$ size and $O(\log n)$ depth for arbitrarily small δ. The success of this strategy is ascribed to an observation we made earlier, that, for any $0 < \varepsilon < \frac{1}{2}$ and $0 < \delta < 1$, an (ε, δ)-nonblocking n-network is an (ε, δ)-rearrangeable n-network, and an (ε, δ)-rearrangeable n-network is an (ε, δ)-n-superconcentrator. Thus a lower bound (for size or depth) of the (ε, δ)-n-superconcentrator is a lower bound of all three, and an upper bound of the (ε, δ)-nonblocking n-network is an upper bound of all three.

The lower bounds are proved in §5. In §6 we construct the (ε, δ)-nonblocking n-network. A few observations on our upper bound are in order. First, the upper bound is based on an explicit construction and is not merely an existence proof. Second, with high probability we can find a nonblocking network contained in the fault-tolerant network merely by discarding faulty components and their immediate neighbors, so no difficult computations are hidden here. Third, because the contained network is "strictly" nonblocking (see Feldman, Friedman, and Pippenger [FFP] for details), routing can be performed by a "greedy" application of a standard path-finding algorithm, so again no difficult computations are involved.

5. The lower bounds. The strategy of the lower bound proof is as follows. We associate with each input a neighborhood containing all vertices within a logarithmic distance of the input. We show that, for a large set of inputs, these neighborhoods are disjoint (otherwise, two inputs would be shorted by closed failures with high probability). This gives the lower bound for depth. We then partition the vertices in the neighborhoods of these inputs into zones according to their distance from the input. We show that, for a large number of inputs, each of these zones must have logarithmic size (otherwise, some input would be isolated by open failures with high probability). Summing over the zones of each neighborhood and the neighborhoods of the inputs gives the lower bound for size.

Given a graph G, we say the distance from vertex ξ_1 to vertex ξ_2, dist (ξ_1, ξ_2), is the number of edges in the shortest path (not necessarily directed) from ξ_1 to ξ_2; the distance from a vertex ξ to an edge $e = \langle \nu, \mu \rangle$, dist (ξ, e), is min $\{\text{dist}(\xi, \mu), \text{dist}(\xi, \nu)\} + 1$.

LEMMA 1. *A tree with l leaves, in which every internal node has degree at least* 3, *contains at least $l/42$ edge-disjoint paths, each joining* 2 *leaves, and each having length at most* 3.

Proof. We begin by observing that we may assume that each internal node has degree exactly 3. For, if not, we may replace each internal node with degree $d > 3$ by a "tree" comprising $d - 2$ new nodes with degree 3. If we find a set of edge-disjoint paths of length at most 3 in the resulting tree, these will correspond to edge-disjoint paths of no greater length in the original graph. Suppose then that T is a tree with l leaves in which each internal node has degree 3. Clearly, there must be $l - 2$ internal nodes. Let us say that a leaf L is "bad" if there is no other leaf with distance at most 3 from L. We show that there are at most $6l/7$ bad leaves. If L is bad, there are seven internal nodes with distance at most 3 from L (see Fig. 1). Let L "pay" one dollar to each of those nodes. We claim that each of the $l - 2$ internal nodes "collects" at most six dollars, from which it follows that there are at most $6(l - 2)/7 \le 6l/7$ bad leaves. If some internal node V collects more than six dollars from bad leaves at distance at most 3, then more than one of these bad leaves must be adjacent to one of the six or fewer nodes at distance 2 from V. However, no more than one bad leaf can be adjacent to an internal node (see Fig. 2). Thus at least $l/7$ leaves are "good" (that is, not bad). Suppose that there are m good leaves. Let $\mathscr{L}$ be a maximal set of edge-disjoint paths, each joining two good leaves and each having length at most 3. Say that a good leaf is "lucky" if it is the endpoint of a path in $\mathscr{L}$, and that it is "unlucky" otherwise. If L is unlucky, there must be a path P in $\mathscr{L}$ within distance 2 of L. (There is a leaf within distance 3 of L, since L is good, and only a path in $\mathscr{L}$ could prevent L from being joined to such a leaf in the maximal set $\mathscr{L}$.) Let each unlucky leaf "pay" one dollar to some such path P. Each path P "collects" at most four dollars from unlucky leaves, since there are at most four leaves with distance at most 2 from P (see Fig. 3). It follows that there are at most four unlucky leaves for each path in $\mathscr{L}$. Since there are exactly two lucky leaves for each path in $\mathscr{L}$, and $m \ge l/7$ good leaves (lucky and unlucky), this implies that there are at most $m/6 \ge l/42$ paths in $\mathscr{L}$. □

Remark. The bound "$l/42$" in Lemma 1 can be improved to "$l/4$," but this requires a more elaborate analysis, which will be presented elsewhere (see Lin [L]).

COROLLARY 1. *A forest F of l leaves, in which every internal node has degree at least* 3, *contains at least $l/42$ edge-disjoint paths, each joining* 2 *leaves, and each having length at most* 3.

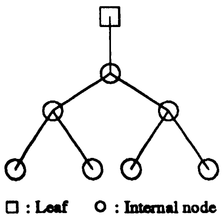


FIG. 1. *A bad leaf.*

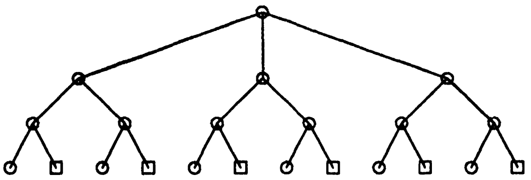


FIG. 2. *An internal node collects at most six dollars.*

LEMMA 2. *Let Φ be a $(\frac{1}{4}, \frac{1}{2})$-$n$-superconcentrator. For all sufficiently large n, at least $n/2$ inputs of Φ have distance (ignoring the direction of each edge) at leasi $(\frac{1}{8}) \log_2 n$ from each other input.*

Proof. Suppose that each of $n/2$ inputs v has a path $\pi(v)$ of length at most j to some other input. We obtain a contradiction if $j = (\frac{1}{8}) \log_2 n$ and n is sufficiently large. Define a forest by (1) starting with the empty forest, (2) considering each such input in turn, and (3) adding to the forest the longest initial segment of $\pi(v)$ that is edge-disjoint from the forest generated thus far. Thus the resulting forest F has at least $n/2$ leaves, and each "stretch" (sequence of consecutive vertices of degree 2) has length at most j. Let G be the forest obtained from F by replacing each stretch, together with the edges incident with its vertices, by a single edge. In G every internal node is of degree at least 3, so we may apply Corollary 1 to obtain at least $n/84$ edge-disjoint paths, each having length at most 3 and each joining one leaf to another. Replacing each edge of these paths by the corresponding stretch, we obtain in F at least $n/84$ edge-disjoint paths, each having length at most $3j$ and each joining one input of Φ to another. Note that, if each of the $3j$ edges on one of the $n/84$ paths is in the closed failure state, two inputs of Φ will contract to a single vertex, and the result will certainly not be an n-superconcentrator. Since this can happen with probability at most $\frac{1}{2}$, we have $1 - (1 - (\frac{1}{4})^{3j})^{n/84} < \frac{1}{2}$. If we set $j = (\frac{1}{6}) \log_2 (n/(84 \ln 2))$, we obtain a contradiction using the inequality $(1-x)^y < e^{-xy}$. Thus, if we set $j = (\frac{1}{8}) \log_2 n$, we obtain a contradiction for all sufficiently large n. □

THEOREM 1. *Let Φ be a $(\frac{1}{4}, \frac{1}{2})$-$n$-superconcentrator. For all sufficiently large n, Φ has size at least $(\frac{1}{256}) n (\log_2 n)^2$ and depth at least $(\frac{1}{16}) \log_2 n$.*

Proof. Say an input is "good" if it has distance at least $(\frac{1}{8}) \log_2 n$ from each other input. By Lemma 2, there are at least $n/2$ good inputs. (Note that the existence of two good inputs implies, by the triangle inequality of the distance, that the depth is at least $(\frac{1}{16}) \log_2 n$.) For each good input v, let $B(v)$ denote the set of all edges at distance at

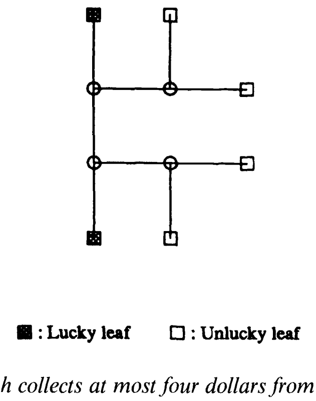


FIG. 3. *Each path collects at most four dollars from unlucky leaves.*

most $(\frac{1}{16})\log_2 n$ from v. For any pair v and w of good inputs, the sets $B(v)$ and $B(w)$ must be disjoint, since otherwise the distance between v and w would be less than $(\frac{1}{8})\log_2 n$. Thus it will suffice to show that, for each good input v, the set $B(v)$ contains at least $(\frac{1}{128})(\log_2 n)^2$ edges for all sufficiently large n. If an input v_0 has all n outputs adjacent to some edges in $B(v_0)$, then it is certainly true that $|B(v_0)| \geq (\frac{1}{128})(\log_2 n)^2$, since the number of edges in $B(v_0)$ cannot be less than the number of outputs adjacent to these edges, and $n \geq (\frac{1}{128})(\log_2 n)^2$ for all sufficiently large n. Thus we may assume that, for each good input v, there is an output $w(v)$ that is not adjacent to an edge in $B(v)$. Consider an arbitrary good input v. Set $i = \lfloor(\frac{1}{16})\log_2 n\rfloor \geq (\frac{1}{32})\log_2 n$. Partition $B(v)$ into subsets $B_1(v), \ldots, B_i(v)$, where $B_h(v)$ comprise those edges at distance h from v. Let $B^*(v)$ denote the set $B_h(v)$ with the minimum number of edges. It will suffice to show that each set $B^*(v)$ contains at least $(\frac{1}{4})\log_2 n$ edges. Let b be the cardinality of the set $B^*(v)$ with the minimum number of edges. It will suffice to show that $b \geq (\frac{1}{4})\log_2 n$ for all sufficiently large n. Consider an arbitrary good input v. Any path from v to $w(v)$ must contain an edge in $B^*(v)$, since the distance from v can increase at most 1 at each successive edge of a path. If edges of $B^*(v)$ are all in open state, all paths from v to $w(v)$ are broken, and the resulting network is certainly not an n-superconcentrator. This can happen with probability at most $\frac{1}{2}$. Thus we have $1 - (1 - (\frac{1}{4})^b)^{n/2} < \frac{1}{2}$. As before, this implies that $b \geq (\frac{1}{2})\log_2 (n/2 \ln 2) \geq (\frac{1}{4})\log_2 n$ for all sufficiently large n. □

6. The upper bounds. In this section, we explicitly construct $(10^{-6}, \delta)$-nonblocking n-networks with $O(n(\log n)^2)$ edges and $O(\log n)$ depth for arbitrarily small δ.

The strategy of the upper bound proof is as follows. We use a standard recursive construction for nonblocking networks, but scale the construction up by a logarithmic factor and terminate the recursion with subnetworks of logarithmic size (rather than constant size). We then use networks (called "directed grids" in this paper) of logarithmic by logarithmic size based on the "hammock" of Moore and Shannon [MS] to interface the inputs and outputs to the terminal subnetworks.

The basic building blocks of the construction are (c, c', t)-*expanding graphs* and (l, w)-*directed grids*. A (c, c', t)-*expanding graph* is a bipartite directed graph with two distinguished sets of t vertices called *inlets* and *outlets*, respectively, where every set of c inlets is joined by edges to at least c' outlets (that is, for every set C of c inlets, there exist a set C' of c' outlets, such that, for every outlet ζ' and C', there is an inlet ζ in C and an edge (ζ, ζ')). The constructions of (an, bn, n)-*expanding graphs* (where $0 < a < b < 1$ are constants) with linear sizes (with respect to n) are quite standard. See Bassalygo and Pinsker [BP] for the probabilistic version, and see Gabber and Galil [GG] for the explicit construction. (We need to mention that the first explicit construction was presented by Margulis [M] and currently the best-known explicit construction is due to Lubotzky, Phillips, and Sarnak [LPS].) An (l, w)-*directed grid* is a directed graph with w stages and l vertices in each stage. A vertex in the jth stage and the ith row is denoted by a binary tuple (i, j), $1 \leq i \leq l$ and $1 \leq j \leq w$. An edge from vertex (i, j) to vertex (i', j') exists if and only if $i' = i$ and $j' = j + 1$ or $i' = i + 1$ and $j' = j + 1$. (See Fig. 4.)

Suppose that we wish to construct an (ε, δ)-nonblocking n-network with $n = 4^\nu$. Set $\gamma = \lceil \log_4 (34\nu) \rceil$, so that $136\nu \geq 4^\gamma \geq 34\nu$. We first construct a nonblocking $4^{\nu+\gamma}$-network through the recursive construction illustrated in Pippenger [P82, §9]. (This network is a directed graph with $2(\nu + \gamma) + 1$ stages, with $4^{\nu+\gamma}$ inputs on stage 0 and $4^{\nu+\gamma}$ outputs on stage $2(\nu + \gamma)$. Each other stage contains $64 \cdot 4^{\nu+\gamma}$ vertices. Edges only exist between some vertices in adjacent stages. The subgraph induced by inputs and vertices in stage 1 consists of $4^{\nu+\gamma-1}$ disjoint bipartite graphs, each having four inputs

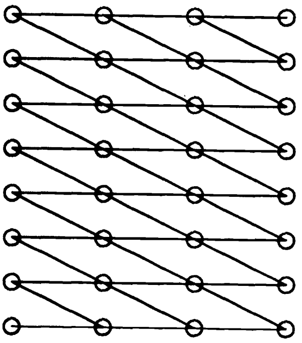


FIG. 4. *A* (4, 8)*-directed grid.*

on one side and 256 vertices on the other side. Similar property holds for the subgraph induced by outputs and vertices in the adjacent stage. The subgraph induced by vertices in stage i and stage $i + 1$ (for all $1 \leq i \leq \nu + \gamma - 1$) consists of $4^{\nu+\gamma-i}$ disjoint $(32 \cdot 4^i, 32(1 + (2 - \sqrt{3})/8) \cdot 4^i, 64 \cdot 4^i)$-expanding graphs, with each vertex on stage i having ten out-edges and vertex on stage $i + 1$ ten in-edges. The subnetwork from stage $\nu + \gamma$ to stage $2(\nu + \gamma)$ is a *mirror image* of that from stage 0 to stage $\nu + \gamma$. Network N' is a mirror image of network N if N' is obtained from N by (1) exchanging the inputs with the outputs and (2) reversing the direction of every edge.) We then remove vertices in the first and last γ stages and edges incident with them and let $\mathcal{M}$ be the remaining graph. The first stage of $\mathcal{M}$ consists of 4^ν disjoint sets vertices, each being the inlets of a $(32 \cdot 4^\gamma, 33.07 \cdot 4^\gamma, 64 \cdot 4^\gamma)$-expanding graph (note that $32(1 + (2 - \sqrt{3})/8) > 33.07$). Construct 4^ν $(\nu, 64 \cdot 4^\gamma)$-directed grids $\Phi_1, \ldots, \Phi_{4^\nu}$. Joined to each vertex in the first stage of Φ_i ($i = 1, \ldots, 4^\nu$) is an edge from an *input* vertex. Combine $\mathcal{M}$ with $\Phi_1, \ldots, \Phi_{4^\nu}$ and the associated inputs by (1) letting the 4^ν $(32 \cdot 4^\gamma, 33.07 \cdot 4^\gamma, 64 \cdot 4^\gamma)$-expanding graphs in the first stage of $\mathcal{M}$ correspond to $\Phi_1, \ldots, \Phi_{4^\nu}$ in any one-to-one fashion, and (2) in each such corresponding pair, identifying the inlets of the expanding graph with the vertices in the last stage of the directed grids in any one-to-one fashion. Similarly, construct 4^ν $(\nu, 64 \cdot 4^\gamma)$-directed grids $\Psi_1, \ldots, \Psi_{4^\nu}$; join an *output* by edges to every vertex in the last stage of each Ψ_j ($j = 1, \ldots, 4^\nu$); combine $\Psi_1, \ldots, \Psi_{4^\nu}$ and the associated outputs with $\mathcal{M}$ (and the above combined $\Phi_1, \ldots, \Phi_{4^\nu}$ and the associated inputs) by identifying vertices of the first stage of $\Psi_1, \ldots, \Psi_{2^\nu}$ with vertices in the last stage of $\mathcal{M}$. Call the resulting network $\mathcal{N}$. (See Fig. 5.)

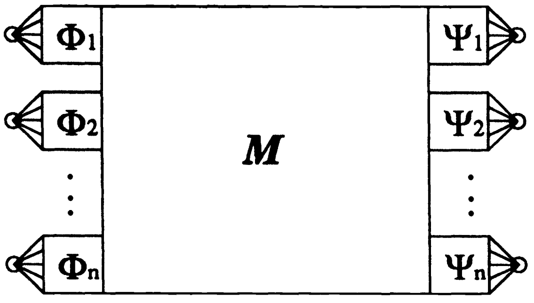


FIG. 5. *Network* $\mathcal{N}$.

Network $\mathscr{N}$ has $2(\nu + \nu) + 1 = 4\nu + 1$ stages. The 4^ν inputs and 4^ν outputs are on stage 0 and stage 4ν, respectively. Each other stage contains $64 \cdot 4^{\nu+\gamma}$ vertices. The subnetwork from stage 1 to stage ν and that from stage 3ν to stage $4\nu - 1$ consist of $\Phi_1, \ldots, \Phi_{4^\nu}$ and $\Psi_1, \ldots, \Psi_{4^\nu}$, respectively. The subnetwork from stage ν to stage 3ν is $\mathscr{M}$. Each input has out-degree $64 \cdot 4^\gamma$; a vertex in Φ_i has in-degree 2 and out-degree 2, except vertices on their first stages (in-degree 1) and last stages (out-degree 10); a vertex in the left-hand half of $\mathscr{M}$ (stage ν to stage 2ν of $\mathscr{N}$) has in-degree 10 and out-degree 10, except vertices on stage ν (in-degree 2). The subnetwork from stage 2ν to stage 4ν (called $\mathscr{N}_{\mathscr{R}}$) is a mirror image of that from stage 0 to stage 2ν (called $\mathscr{N}_{\mathscr{L}}$). In particular, the right-hand half of $\mathscr{M}$, called $\mathscr{M}_{\mathscr{R}}$, is a mirror image of $\mathscr{M}_{\mathscr{L}}$, the left-hand half of $\mathscr{M}$. Network $\mathscr{N}$ has $1408\nu 4^{\nu+\gamma}$ edges because there are $1280\nu 4^{\nu+\gamma}$ edges in $\mathscr{M}$, $128(\nu - 1)4^{\nu+\gamma}$ edges in Φ_i and Ψ_i, for all $1 \le i \le 4^\nu$, and $128 \cdot 4^{\nu+\gamma}$ edges adjacent to inputs and outputs.

Let η be a vertex of $\mathscr{N}$ that is not an input or an output. Say a vertex η of $\mathscr{N}$ is *faulty*, if an edge $\langle \tau, \eta \rangle$ or $\langle \eta, \xi \rangle$ is in open failure or closed failure state. Given a set of vertex-disjoint direct paths from inputs to outputs in $\mathscr{N}$, for an input, an output, or a vertex that is not faulty, it is said to be *idle* if it is not involved in these paths, *busy* otherwise. Say an (idle) vertex ξ_1 has *access* to another (idle) vertex ξ_2 if there is a path of idle vertices from ξ_1 to ξ_2. It is clear that, if ξ_1 has access to ξ_2 and ξ_2 has access to ξ_3, then ξ_1 has access to ξ_3. A network N is a *majority-access* network if, given any set of directed paths from inputs to outputs, every idle input has access to a majority (strictly more than half) of the outputs.

LEMMA 3. *Let ξ be an idle input of network $\mathscr{N}$. The probability that ξ has access to at least $32 \cdot 4^\gamma + 1$ vertices in the last stage of Φ_ξ (i.e., strictly more than half) is at least $1 - c_1\nu(144\varepsilon)^\nu$, where $c_1 = 1/(1 - 72\varepsilon)$.*

Proof. Let us begin by estimating the probability that ξ does not have access to any vertex at the last stage of Φ_ξ. There is no busy vertex in Φ_ξ, since ξ is idle and $\mathscr{N}$ is a directed and staged graph. By Menger's theorem (see, e.g., Chapter 5 of [CL]), there is a "(vertex) cut set" of Φ_ξ (i.e., the removal of which and their adjacent edges will separate ξ and the vertices at the last stage) consisting of faulty vertices only. Consider a cut set C of l vertices. Then it must be $l \ge 64 \cdot 4^\gamma$, since Φ_ξ has this many rows. If every vertex in C is faulty, the probability is at most $(24\varepsilon)^l$, since each vertex in Φ_ξ (other than ξ, which is not in any cut set we consider) is adjacent to at most twelve edges. For any given l, the number of such cut set C is at most $\nu 3^l$, since they are ν vertices at the first row of Φ_ξ to start C, and at most three ways (each along an edge) to continue at each step. Thus the probability that ξ does not have access to any vertex at the last stage of Φ_ξ is at most

$$\sum_{l \ge 64 \cdot 4^\gamma} \gamma 3^l (24\varepsilon)^l = c_1\nu(72\varepsilon)^{64 \cdot 4^\gamma}.$$

Consider an arbitrary set S of $32 \cdot 4^\gamma$ vertices in the last stage of Φ_ξ. The probability that ξ does not have access to any vertex in S is at most $c_1\nu(72\varepsilon)^{64 \cdot 4^\gamma}$. There are at most

$$\binom{64 \cdot 4^\gamma}{32 \cdot 4^\gamma} < 2^{64 \cdot 4^\gamma}$$

such S. This implies that the probability of ξ having access to at least $32 \cdot 4^\gamma + 1$ vertices in the last stage of Φ_ξ is at least $1 - c_1\nu(144\varepsilon)^\nu$, since $64 \cdot 4^\gamma > \nu$. □

LEMMA 4. *In a $(32 \cdot 4^\mu, 33.07 \cdot 4^\mu, 64 \cdot 4^\mu)$-expanding graph in $\mathscr{M}_{\mathscr{L}}$, for any $\gamma \le \mu \le \nu + \gamma - 1$ (the expanding graph is in the subgraph from stage $\mu + \nu - \gamma$ to stage $\mu + \nu - \gamma + 1$ of $\mathscr{N}$), the probability that it has more than $0.07 \cdot 4^\mu$ outlets faulty is at most $e^{-0.06 \cdot 4^\mu}$.*

Proof. There are $1280 \cdot 4^\mu$ edges incident with outlets of the $(32 \cdot 4^\mu, 33.07 \cdot 4^\mu, 64 \cdot 4^\mu)$-expanding graph (each vertex has ten in-edges and ten out-edges). For each such edge, let x_j be the random variable such that $x_j = 0$ if the edge is in normal state, $x_j = 0$ otherwise, for all $1 \le j \le 1280 \cdot 4^\mu$. It is clear that $\Pr[x_j = 1] < 2\varepsilon$ and $\Pr[x_j = 0] > 1 - 2\varepsilon$. Let $T = \sum_{j=1}^{1280 \cdot 4^\mu} x_j$,

$$\Pr[T > 0.07 \cdot 4^\mu] = \Pr[e^T > e^{0.07 \cdot 4^\mu}] < \mathrm{E}[e^T]/e^{0.07 \cdot 4^\mu}$$

by Markov's inequality. As x_j's are independent,

$$\mathrm{E}[e^T] = \prod_{j=1}^{1280 \cdot 4^\mu} \mathrm{E}[e^{x_j}] < (1 + 2\varepsilon e)^{1280 \cdot 4^\mu} < e^{2560 e\varepsilon \cdot 4^\mu},$$

since $(1 + x)^y < e^{xy}$, and $2560 e\varepsilon < 0.01$ when $\varepsilon = 10^{-6}$. Thus the probability that there are more than $0.07 \cdot 2^\mu$ outlets faulty is at most $e^{0.01 \cdot 4^\mu - 0.07 \cdot 4^\mu} = e^{-0.06 \cdot 4^\mu}$. □

LEMMA 5. *The probability that there exists a* $(32 \cdot 4^\mu, 33.07 \cdot 4^\mu, 64 \cdot 4^\mu)$-*expanding graph in* $\mathcal{M}_{\mathcal{L}}$ *with more than* $0.07 \cdot 4^\mu$ *faulty outlets, for some* $\gamma \le \mu \le \nu + \gamma - 1$, *is less than* $\nu(2/e)^{2\nu}$.

Proof. It is simply a problem of counting the number of expanding graphs with respect to the number of outlets. There are $4^{\nu+\gamma-\mu}$ $(32 \cdot 4^\mu, 33.07 \cdot 4^\mu, 64 \cdot 4^\mu)$-expanding graphs between stage $\nu + \mu - \gamma$ and stage $\nu + \mu - \gamma + 1$ of $\mathcal{M}_{\mathcal{L}}$, for all $\gamma \le \mu \le \nu + \gamma - 1$. By Lemma 4, the probability that there is an expanding graph with no more than $0.07 \cdot 4^\mu$ outlets is at most

$$\sum_{\mu=\gamma}^{\nu+\gamma-1} 4^{\nu+\gamma-\mu} e^{-0.06 \cdot 4^\mu} < \sum_{\mu=\gamma}^{\nu+\gamma-1} 4^\nu e^{-0.06 \cdot 4^\gamma} < \nu 4^\nu e^{-2\nu},$$

since $4^\gamma \ge 34\nu$. □

LEMMA 6. *With probability at least* $1 - c_1\nu(144\varepsilon)^\nu - \nu(2/e)^{2\nu}$, $\mathcal{N}_{\mathcal{L}}$ *is a majority-access network.*

Proof. We may assume in this proof that each $(32 \cdot 4^\mu, 33.07 \cdot 4^\mu, 64 \cdot 4^\mu)$-expanding graph in $\mathcal{M}_{\mathcal{L}}$ has at least $0.07 \cdot 4^\mu$ outlets faulty, and each idle input ξ has access to at least $32 \cdot 4^\gamma + 1$ vertices in the last stage of Φ_ξ. It is clear by Lemmas 3 and 5 that the probability of the assumption failing is at most $1 - c_1\gamma(144\varepsilon)^{64 \cdot 4^\gamma} - \nu(2/e)^{2\nu}$. For each pair of inputs ξ_1 and ξ_2, we say their relativity, relat (ξ_1, ξ_2) = relat (ξ_2, ξ_1), is d $(1 \le d \le \nu)$ if and only if the directed paths starting from the two inputs may share a vertex at or after the $(d + \nu)$th stage but cannot share any vertex before the $(d + \nu)$th stage. It is observed that, for each input ξ, there are $4^d - 4^{d-1} = 3 \cdot 4^{d-1}$ other inputs ξ' with relat $(\xi, \xi') = d$, for any d with $1 \le d \le \nu$. Now suppose that ξ is an arbitrary idle input, let the subnetwork N_0 of $\mathcal{N}$ be Φ_ξ, and let N_k be the subnetwork induced by vertices that can only be reached by ξ and $4^k - 1$ other inputs ξ' with relat $(\xi, \xi') \le k$. It is clear that N_k has 4^k inputs and $64 \cdot 4^{\gamma+k}$ outputs. We prove by induction on k that ξ has access to at least $32 \cdot 4^{\gamma+k} + 1$ outputs of N_k, thus, in particular, has access to strictly more than half of the outputs of $\mathcal{N}_{\mathcal{L}} = N_\nu$. The base case N_0 is obviously true because of our assumption. Consider N_{k+1}. The outputs of N_k are linked to the output of N_{k+1} via four $(32 \cdot 4^{\gamma+k}, 33.07 \cdot 4^{\gamma+k}, 64 \cdot 4^{\gamma+k})$-expanding graphs (with the inlets being the outputs of N_k and the outlets being four disjoint subsets of the outputs of N_{k+1}). By the induction hypothesis, ξ has access to at least $32 \cdot 4^{\gamma+k} + 1$ outputs of N_k. These $32 \cdot 4^{\gamma+k} + 1$ vertices are joined by edges to at least $4 \cdot 33.07 \cdot 4^{\gamma+k}$ outputs of N_{k+1} (via the four $(32 \cdot 4^{\gamma+k}, 33.07 \cdot 4^{\gamma+k}, 64 \cdot 4^{\gamma+k})$-expanding graphs). By our assumption, there are at most $0.07 \cdot 4^{\gamma+k+1}$ outputs of N_{k+1} are faulty. There are at most $4^{k+1} - 1$ outputs of N_{k+1} that are busy, because each busy output is one-to-one corresponding (via a directed path) to a busy input of N_{k+1}, and there are at most $4^{k+1} - 1$ such inputs. Thus, ξ has access to at least $4 \cdot 33.07 \cdot 4^{\gamma+k} - 0.07 \cdot 4^{\gamma+k+1} - 4^{k+1} + 1 > 32 \cdot 4^{\gamma+k+1} + 1$ outputs of N_{k+1}. This completes our induction. □

COROLLARY 2. *With probability at least* $1 - c_1\nu(144\varepsilon)^{\nu} - \nu(2/e)^{2\nu}$, *the mirror image of* $\mathcal{N}_{\mathcal{R}}$ *is a majority-access network.*

We observe that, if $\mathcal{N}_{\mathcal{L}}$ and the mirror image of $\mathcal{N}_{\mathcal{R}}$ are both majority-access networks and the inputs and outputs of $\mathcal{N}$ are distinct (no two input(s) and output(s) contracting to a single vertex), then $\mathcal{N}$ contains a nonblocking 4^{ν}-network of no-failure edges.

LEMMA 7. *With probability at most* $c_2\nu^2(160\varepsilon)^{2\nu}$, *where* $c_2 = 4^{15}/(1 - 40\varepsilon)$, *there exist two input(s) and output(s) that contract to a single vertex.*

Proof. The correctness of the lemma follows four observations. First, any simple path joining two input(s) and output(s) must contain at least 2ν edges. Second, for any $l \geq 2\nu$, there are at most $(64\cdot 4^{\gamma})^2(40)^{l-2}$ such paths of length l, since the degree of inputs and outputs is $64 \cdot 4^{\gamma}$ and that of the other vertices is at most 40. Note that $(64\cdot 4^{\gamma})^2(40)^{l-2} < 4^{14}\nu^2(40)^l$, since $4^{\gamma} \leq 136\nu$. Third, the probability that a path of length l gets "shorted" (all edges on the path are in closed failure state) is less than ε^l. Last, there are at most $(2\cdot 4^{\nu})^2$ such input or output pairs. □

THEOREM 2. *Network* $\mathcal{N}$ *is a* $(10^{-6}, \delta)$*-nonblocking n-network with at most* $4^9 n(\log_4 n)^2$ *edges and* $5 \log_4 n$ *depth for arbitrarily small* δ, *when n is sufficiently large.*

Proof. We have seen that network $\mathcal{N}$ contains at most $1408\nu 4^{\nu+\gamma}$ edges and has $4\nu + 1$ depth, where $n = 4^{\nu}$ and $\gamma = \lceil \log_4 34\nu \rceil$. Work out the constant using $4^{\gamma} \leq 136\nu$. The probability that $\mathcal{N}$ fails to contain a nonblocking n-network of no-failure edges is less than $2(c_1\nu(144\varepsilon)^{\nu} - \nu(2/e)^{2\nu}) + c_2\nu^2(160\varepsilon)^{2\nu}$, by Lemma 6, Corollary 2, and Lemma 7. This value can be arbitrarily small when $n = 4^{\nu}$ is sufficiently large, given $\varepsilon = 10^{-6}$. □

Acknowledgment. The authors thank Bruce Maggs for helpful discussions.

REFERENCES

[AHU] A. V. AHO, J. E. HOPCROFT, AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[ALM] S. ARORA, T. LEIGHTON, AND B. MAGGS, *On-line algorithms for path selection in a nonblocking network*, ACM Sympos. on Theory of Computing, 22 (1990), pp. 149–158.

[B] V. E. BENEŠ, *Optimal rearrangeable multistage connecting networks*, Bell System Tech. J., 43 (1964), pp. 1641–1656.

[BP] L. A. BASSALYGO AND M. S. PINSKER, *Complexity of optimum non-blocking switching network without reconnections*, Problems Inform. Transmission, 9 (1974), pp. 64–66.

[Cl] C. CLOS, *A study of non-blocking networks*, Bell System Tech. J., 32 (1953), pp. 406–424.

[Co] M. I. COLE, *Algorithmic Skeletons: A Structured Approach to the Management of Parallel Computation*, Ph.D. thesis, Computer Science, Univ. of Edinburgh, Oct. 1988.

[CL] G. CHARTRAND AND L. LESNIAK, *Graphs and Digraphs*, 2nd ed., Wadsworth, Belmont, CA, 1986.

[FFP] P. FELDMAN, J. FRIEDMAN, AND N. PIPPENGER, *Wide-sense non-blocking networks*, SIAM J. Discrete Math., 1 (1988), pp. 158–173.

[GG] O. GABBER AND Z. GALIL, *Explicit constructions of linear-sized superconcentrators*, J. Comput. System Sci., 22 (1981), pp. 407–420.

[L] G. LIN, *Edge-disjoint paths in a tree*, in preparation.

[L92] ———, *Fault-tolerant planar communication networks*, ACM Sympos. on Theory of Computing, 24 (1992), pp. 133–139.

[LM] T. LEIGHTON AND B. MAGGS, *Expanders might be practical: Fast algorithms for routing around faults on multibutterflies*, IEEE Sympos. on Foundation of Computer Science, 30 (1989), pp. 384–389.

[LPS] A. LUBOTZKY, R. PHILLIPS, AND P. SARNAK, *Ramanujan graphs*, Combinatorica, 8 (1988), pp. 261–277.

[M] G. A. Margulis, *Explicit constructions of concentrators*, Problems Inform. Transmission, 9 (1975), pp. 325–332. (In English.)

[MS] E. F. Moore and C. E. Shannon, *Reliable circuits using less reliable relays*, Part I and Part II, J. Franklin Inst., 262 (1956), pp. 191–208, 281–297.

[PY] N. Pippenger and A. C. Yao, *Rearrangeable networks with limited depth*, SIAM J. Algebraic Discrete Math., 3 (1982), pp. 411–417.

[P82] N. Pippenger, *Telephone switching networks*, AMS Proc. Sympos. Appl. Math., 26 (1978), pp. 101–133.

[P90] ———, *Communication networks*, in Handbook of Theoretical Computer Science, Chapter 15, J. van Leeuwen, ed., Elsevier, Amsterdam, 1990.

[S] C. E. Shannon, *Memory requirements in a telephone exchange*, Bell System Tech. J., 29 (1950), pp. 343–349.

[U] E. Upfal, *An $O(\log N)$ deterministic packet routing scheme*, ACM Sympos. on Theory of Computing, 21 (1989), pp. 241–250.

[V] L. G. Valiant, *On nonlinear lower bounds in computational complexity*, ACM Sympos. on Theory of Computing, 7 (1975), pp. 45–53.

SIAM J. DISCRETE MATH.
Vol. 7, No. 1, pp. 119–132, February 1994

012

SORTING ON A MESH-CONNECTED COMPUTER WITH DELAYING LINKS*

BOGDAN S. CHLEBUS†, KRZYSZTOF DIKS†, AND ANDRZEJ PELC‡

Abstract. A mesh-connected processor array is considered in which the links are faulty in the following sense: Each attempt by two neighboring processors to communicate by exchanging messages may fail with some constant probability. A message sent across a link and not delivered is said to be delayed by the link. It is assumed that all the links delay with the same fixed delay probability, independently of each other.

The problem of sorting is addressed in this model. It is proved that an $n \times n$ mesh can be sorted in the expected time $O(n)$ with large probability. More precisely, it is shown that there are two constants $c > 0$ and $r > 1$, depending on the delay probability, such that the $n \times n$ mesh is sorted in time $cn + t$ with the probability at least $1 - r^{-t}$. One specific algorithm is considered, but the analysis shows that many known algorithms could sort in the expected time $O(n)$, after some natural modifications.

Key words. parallel sorting, mesh-connected computer, fault tolerance

AMS subject classifications. 68M, 68M10, 68M15, 68Q25, 60Y20

1. Introduction. Parallel computers with a large number of components may have some of the processors or links not fully operational. Hence it is an important feature of algorithms designed for such computers to be efficient in the presence of faults.

Faulty components can be introduced to a model of communication network in many ways. The most-often-investigated scenario assumes strictly that, if a processor or link is faulty, then it does not perform any of the specified set of functions; next, it is stipulated whether faults are generated randomly or if the worst-case is considered. A communication network with such *permanently faulty* components might be tried to be reconfigured to obtain a network of the same pattern of interconnections among processors but of a smaller size. Hastad, Leighton, and Newman [4] showed how to reconfigure hypercube networks with random faulty components, and Kaklamanis et al. [5] studied faulty arrays of processors and their reconfigurations. These results have shown that both hypercubes and meshes are very robust in the presence of randomly generated (permanently) faulty components.

In this paper, we study a different model of a network with faulty components. We assume that all the processors are operational but the links may *delay* messages. This means that an attempt to exchange messages along a link may fail with some fixed probability p. This number p is called the *delay probability*. If a message is not delivered, then the processors may keep repeating attempts to communicate until eventually there is a success. It is assumed that all the links delay with the same probability independently of each other.

A network with delaying links does not need to be reconfigured. What becomes a problem is to design algorithms that are robust in this situation and to evaluate their expected behavior. In this paper, we consider $n \times n$ meshes of processors with delaying links and show that they can sort in time $O(n)$ with large probability. The analysis shows that many known algorithms can be readily adapted to this model and sort within time $O(n)$.

*Received by the editors April 22, 1991; accepted for publication (in revised form) October 2, 1992.

†Instytut Informatyki, Uniwersytet Warszawski, Banacha 2, 02-097 Warszawa, Poland. The research of this author was supported in part by grants KBN 2-1190-91-01, KBN 2-2044-92-03, and KBN 2-2043-92-03.

‡ Département d'Informatique, Université du Québec à Hull, C.P. 1250, succ. "B," Hull, Québec J8X 3X7, Canada. This research was supported in part by Natural Sciences and Engineering Research Council of Canada grant OGP 0008136.

A large part of the analysis described in this paper is devoted to the odd-even transposition sort running on a line of processors. If links are delaying, then the behavior of the algorithm is shown to be bound by the behavior of the following stochastic process. There is a binary tree with a series of queues associated with each node. The number of queues at a node depends on the depth of the node. There are a number of customers, and each of them must be served by all the nodes on some branch. The customers go through all the queues at a node and next are redirected to the children. The process terminates when all customers have been served.

The paper is organized as follows. Section 2 contains the analysis of the odd-even transposition sort and is broken into sections about the series of queues, the tree process, and the main theorem. Section 3 is about the analysis of a sorting algorithm on the square meshes. Section 4 contains remarks and open problems.

2. Line of processors. In this section, we study the odd-even transposition sort on a line of processors connected by delaying links. It is shown that the algorithm completes sorting in time proportional to the length of the line, with a large probability. The last section contains a theorem giving estimates on this probability. This theorem can be also derived from the results of Berman and Simon [1]. The probabilistic analysis of this section was applied by the authors to evaluate the behavior of a gossiping protocol in a faulty environment in [2].

The odd-even transposition sort operates as follows: All the odd-numbered processors exchange their keys with the right neighbors in the odd-numbered steps and with the left neighbors in the even-numbered steps. If an exchange is successful, then the left processor keeps the smaller key and the right one the larger key. It is known that with all the links being operational it takes at most n steps to sort the line.

An often-used approach to proving correctness of a sorting algorithm is by using the $0-1$ principle, which states that the algorithm sorts all inputs if it sorts the inputs of 0's and 1's. This principle is valid if the algorithm sorts by conditional exchanges, because we can interpret 0's as the keys smaller than some specific key and 1's as the remaining keys. Our approach is similar, the difference being that the keys dividing the sequence into 0's and 1's are directly specified during various stages of the algorithm.

The odd-even transposition sort can be conceptually divided into the following stages. Partition all the keys into two equal groups A and B, where each element of A is less than or equal to each element of B. Wait until all the (smaller) elements of A are to the left of all the (greater) elements of B. Then apply the same procedure to two halves: Divide them into equal groups of smaller and larger keys and wait until the smaller are to the left of the larger. Repeat this iteratively until the whole line is sorted. In the following, the elements of A are interpreted as 0's and the elements of B as 1's. The whole process can be visualized as a tree. There is a line of n processors at the root, each storing either a 0 or a 1, the numbers of 0's and 1's being equal. In general, a node with a line of k processors has two children, each with $k/2$ processors and with equal numbers of 0's and 1's. The process is started at the root and terminates when all the nodes are done. Starting a node means starting to sort the associated line of processors. As soon as a node becomes sorted, then both of its children are started. The time is the maximum over all the branches of the tree. We assume that n is a power of 2.

Consider a node of the tree and a line of k processors associated with it. We may assume that initially all the 0's are to the right of the 1's. Our approach is to interpret the 1's as servers and 0's as customers in a queuing system. Each customer must go through all the servers. Once a customer has been served by a server, it is moved to the queue (maybe empty) of the customers waiting for service from the next server. For a 0 (customer) to be served by a 1 (server) means to be exchanged with it. All the 0's between two consecutive 1's are in the same queue waiting for service from the left 1. If two consecutive 1's are adjacent, then this means that the queue is empty.

We must have two things done: One is the analysis of a sequence of queues, each with a geometric time of service, and the second is the analysis of the behavior of a tree that has a series of queues at each node. These are done in the next sections.

2.1. Series of queues. The starting point is a series of n servers with n customers waiting in queue for the first server. Each server processes the customers and in this way generates them for the next server, if there is any. The output of the first server is geometric. The second server operates in a similar way, but, since it must wait for the first customer for at least one unit of time, its output does not have a geometric distribution on the waiting time. The same is true for the remaining servers, who must wait for their first customers even longer. When the system operates long enough, the distribution of the output of each server converges to the geometric one, and the effect of the time to wait for the first customers diminishes. We do not estimate the rate of this convergence; instead, we work directly with the limit stationary distribution. This is done in two steps: First, we assume that the customers are not given but must be generated and, next, that there are some additional "dummy" customers in the system that must be cleared before processing the true customers.

Suppose that, instead of exactly n customers waiting for service from the first server, the customers have been generated with a geometric distribution and, next, that each of them is served by a series of n servers. This system can be interpreted as a (homogeneous) Markov chain where a state is determined by the numbers of customers in each queue. Initially, each queue is empty. Let us first consider a single server.

A single server is also a Markov chain. We consider a situation in which customers are generated with a geometric distribution and they are also served with a geometric distribution, but these two distributions are different. The reason for this is that we need an ergodic chain. A helpful visualization of the situation is obtained by interpreting the process as a discrete random walk in one dimension with a reflecting barrier. There are three possibilities in each step. If a new customer is generated and none completes its service, then the number of customers increases; this is interpreted as a move away from the barrier. Alternatively, a customer completes the service but no new one is generated so the number of customers decreases; this corresponds to a move toward the barrier. In any other situation, the number of customers does not change, and there is no move. It is well known that such a Markov chain is ergodic if and only if the probability of a move toward the barrier is greater than of a move away from it (cf. [6]).

Suppose that in each unit of time there is a new customer generated with some *input probability* p and also, if the queue is nonempty, then the first customer is served with some *output probability* q. So the input and output (if the queue is nonempty) have geometric distributions on their waiting times with parameters p and q, respectively. The system is said to be in state i if there are i customers in the queue. If there are $k > 0$ customers waiting for service, then the probability of decreasing of this number is $l = (1-p)q$, and the probability of increasing it is $r = (1-q)p$. The matrix $\mathrm{P} = (p_{ij})$ of transition probabilities, where

$$p_{ij} = Pr\text{ (state } j \text{ is entered next} \mid i \text{ is the current state}),$$

is as follows:

1. $p_{ij} = 0$ if $|i-j| > 1$;
2. $p_{00} = 1-p,\ p_{10} = l$;
3. $p_{01} = p,\ p_{11} = 1-l-r,\ p_{21} = l$;
4. If $k > 1$, then $p_{k-1,k} = r,\ p_{k,k} = 1-l-r,\ p_{k+1,k} = l$.

This defines an irreducible and aperiodic Markov chain. If $p < q$, then it is ergodic, in the sense that the probability distribution

$$\mathrm{P}^{(k)} = \mathrm{P}^{(0)}\mathrm{P}^{k}$$

converges to some distribution P independently of the input distribution $\mathrm{P}^{(0)}$. The ergodicity follows from the fact that an aperiodic and irreducible Markov chain is ergodic if there is a nonnull solution $x = (x_i)$ of the equation $x = x\mathrm{P}$, with $\sum |x_i| < \infty$ (cf. [6]). There is a unique solution satisfying the conditions $\sum_{i=0}^{\infty} x_i = 1$ and $x_i \geq 0$. It gives the stationary distribution having the property that, if it is the initial distribution $\mathrm{P}^{(0)}$, then all the subsequent distributions $\mathrm{P}^{(n)} = \mathrm{P}^{(0)} \cdot \mathrm{P}^{n}$ are the same as $\mathrm{P}^{(0)}$. For a proof of the next lemma, see Kemeny, Snell, and Knapp [6].

LEMMA 2.1. *If the input probability p is smaller than the output probability q, then the single-queue system is ergodic.*

LEMMA 2.2. *If the initial distribution is equal to the stationary distribution, then the output has a geometric distribution with the parameter being the input probability p.*

Proof. A new customer completes the service with probability q, provided that the queue is nonempty. Hence the unconditional probability of a new customer being served is $(1-x_0)\cdot q$. To find x_0, note that

$$1 = \sum_{i=0}^{\infty} x_i = x_0 + x_1 \sum_{i=0}^{\infty} \left(\frac{r}{l}\right)^i = x_0 + x_1 \cdot \frac{1}{1-\frac{r}{l}}$$

$$= x_0 + x_0 \cdot \frac{p}{l} \cdot \frac{l}{l-r} = x_0 \cdot \frac{l-r+p}{l-r}.$$

Hence

$$x_0 = \frac{l-r}{l-r+p} = 1 - \frac{p}{l-r+p} = \frac{q-p}{q} = 1 - \frac{p}{q},$$

and therefore $(1-x_0)\cdot q = p/q \cdot q = p$. □

A similar phenomenon, as described in the above lemma, occurs if the queue has an exponential service time, this being a continuous-time distribution (cf. Gross and Harris [3]).

Now let us return to the original series of queues. Suppose that, if a queue is nonempty, then the first customer is served with probability q. Suppose also that the input for the first server is generated with the geometric distribution with parameter p, where $0 < p < q$. It follows from Lemma 2 that, if the initial distribution of each queue is the stationary distribution, then the whole series of queues is in a stationary distribution, and the output of the system is described by the geometric distribution with parameter p. This observation is a basis of our approach. We must get n customers through a series of n queues, but the time to accomplish this is even longer if in the beginning there are some d "dummy" customers in the system, where the number d is determined by the stationary distribution. In such a situation, the time to process n "true" customers becomes the sum of the time needed to clear the system of the dummy customers and the time to process the next n customers. First, we estimate the number of the dummy customers.

LEMMA 2.3. *Let s_k be the probability that initially there are k dummy customers in all the n queues. If p is selected, so that $p = q^2$, then the following inequality holds*:

$$s_k \le \binom{n+k-1}{k}\left(\frac{q}{1+q}\right)^k .$$

Proof. The generating function of the geometric distribution with parameter p is

$$Q(x) = p + p(1-p)x + p(1-p)^2x^2 + \cdots,$$

and the sum of n random variables X_i, each having a geometric distribution with parameter p, has the generating function

$$Q^n(x) = \sum_{k=0}^{\infty} \binom{n+k-1}{k} p^n(1-p)^k x^k$$

describing the negative binomial distribution. The stationary distribution of the number of customers in a single queue is not exactly geometric since

$$x_0 = 1 - \frac{p}{q} \quad \text{and} \quad x_i = \left(1 - \frac{p}{q}\right) \cdot \frac{p}{l} \cdot \left(\frac{r}{l}\right)^{i-1} \quad \text{for } i \ge 1.$$

Approximate this sequence by the one with values

$$y_i = a \cdot \left(1 - \frac{r}{l}\right) \cdot \left(\frac{r}{l}\right)^i \quad \text{for } i = 0, 1, 2, \ldots$$

in the sense that the inequality $x_i \le y_i$ should hold for each $i \ge 0$. If $i \ge 1$, then this means that

$$a \cdot \left(1 - \frac{r}{l}\right) \cdot \left(\frac{r}{l}\right)^i \ge \left(1 - \frac{p}{q}\right) \cdot \frac{p}{l} \cdot \left(\frac{r}{l}\right)^{i-1},$$

which is equivalent to $a \ge (1-p)/(1-q)$. The number $a = (1-p)/(1-q)$ is also good for $i = 0$ because of the inequality

$$\frac{1-p}{1-q} \cdot \left(1 - \frac{r}{l}\right) \ge 1 - \frac{p}{q}.$$

Let $P(x) = \sum_{k=0}^{\infty} x_k x^k$ be the generating function of the stationary distribution and let $R(x) = \sum_{k=0}^{\infty} r_k x^k$ be the generating function of the geometric distribution $r_k = (1 - r/l)(r/l)^k$. If $P^n(x) = \sum_{k=0}^{\infty} s_k x^k$ and $[aR(x)]^n = \sum_{k=0}^{\infty} t_k x^k$, then the inequality $s_k \le t_k$ holds because $x_k \le a r_k$. The number s_k is the probability that there are k customers in all the queues, and the number t_k is given by the formula

$$t_k = a^n \binom{n+k-1}{k} \left(1 - \frac{r}{l}\right)^n \cdot \left(\frac{r}{l}\right)^k .$$

This formula can be simplified, since

$$a^n \left(1 - \frac{r}{l}\right)^n = \left[\frac{1-p}{1-q}\left(1 - \frac{p(1-q)}{q(1-p)}\right)\right]^n = \left[\frac{q-p}{(1-q)q}\right]^n .$$

The only requirement on p is the inequality $0 < p < q$ to be true. Pick $p = q^2$ so that $(q-p)/(1-q)q = 1$. Then the bound on s_k becomes

$$s_k \le \binom{n+k-1}{k}\left(\frac{r}{l}\right)^k = \binom{n+k-1}{k}\left(\frac{q}{1+q}\right)^k. \qquad \square$$

Let $S = S_n$ be the random variable equal to the time that it takes to get n true customers served by the series of n queues. Let A_d be the event that initially there are d dummy customers in the system. Then the following equality holds:

$$Pr(S=t) = \sum_{d=0}^{\infty} Pr(S=t|A_d)\cdot Pr(A_d).$$

The distribution of A_d has just been estimated, we need a bound on the number $Pr(S = t|A_d)$. This probability is the same as that of an event when there is a sequence of t Bernoulli trials, each with the probability p of success, and the last trial exactly contributes the $(d+n)$th success. (This follows from Lemma 2.2 and the subsequent discussion.) Hence

$$Pr(S=t|A_d) = \binom{t-1}{d+n-1} p^{d+n}(1-p)^{t-n-d}.$$

Therefore the probability $Pr(S = t)$ is bounded as follows:

$$\begin{aligned} Pr(S=t) &\le \sum_{d=0}^{\infty}\binom{t-1}{d+n-1} p^{d+n}(1-p)^{t-n-d}\binom{n+d-1}{d}\left(\frac{q}{1+q}\right)^d \\ &= p^n(1-p)^{t-n}\sum_{d=0}^{t-n}\binom{t-1}{d+n-1}\binom{n+d-1}{d}\left(\frac{q}{1+q}\cdot\frac{p}{1-p}\right)^d. \end{aligned}$$

Transform the product of two binomial coefficients according to the formula

$$\binom{a}{b}\binom{b}{c} = \binom{a}{c}\binom{a-c}{b-c}$$

and introduce the notation $\lambda = q/(1+q)\cdot p/(1-p)$. This yields

$$Pr(S=t) \le (1-p)^t\left(\frac{p}{1-p}\right)^n \sum_{d=0}^{t-n}\binom{t-1}{d}\binom{t-1-d}{n-1}\lambda^d.$$

The sum can be estimated as follows:

$$\begin{aligned} \sum_{d=0}^{t-n}\binom{t-1}{d}\binom{t-1-d}{n-1}\lambda^d &\le \binom{t-1}{n-1}\sum_{d=0}^{t-n}\binom{t-1}{d}\lambda^d \\ &\le \binom{t-1}{n-1}(1+\lambda)^{t-1}. \end{aligned}$$

Therefore

$$Pr(S=t) \leq [(1-p)(1+\lambda)]^t \begin{pmatrix} t-1 \\ n-1 \end{pmatrix} \left(\frac{p}{1-p}\right)^n \cdot \frac{1}{1+\lambda}.$$

Observe that $(1-p)(1+\lambda) = 1-p+pq/(1+q) < 1$. Estimate the binomial coefficient by the following general inequality:

$$\begin{pmatrix} x \\ y \end{pmatrix} \leq \left(\frac{ex}{y}\right)^y$$

and introduce the notation $\gamma = (1-p)(1+\lambda)$ to obtain

$$Pr(S=t) \leq \gamma^t \left[\frac{(t-1)e}{n-1}\right]^{n-1} \cdot \left(\frac{p}{1-p}\right)^n \cdot \frac{1}{1+\lambda} = \gamma^{t-1} \left[\frac{t-1}{n-1} \cdot \frac{ep}{1-p}\right]^{n-1} p.$$

LEMMA 2.4. *There are two constants* d *and* c, *where* $0 < d < 1$ *and* $c > 0$, *such that*

$$Pr(S=t) \leq (1-d)d^t,$$

for $t \geq cn$.

Proof. Take as d any number satisfying the inequalities $\gamma < d < 1$. It follows from the above considerations that, to have the inequality $Pr(S=t) \leq (1-d)d^t$, it is sufficient to have

$$(1) \qquad p\gamma^{t-1} \left(\frac{t-1}{n-1} \cdot \frac{ep}{1-p}\right)^{n-1} \leq (1-d)d^t.$$

Denote

$$b = \frac{ep}{1-p}, \quad f = \frac{p}{(1-d)d}, \quad \text{and} \quad \xi = \frac{\gamma}{d}.$$

Then inequality (1) can be rewritten as

$$(2) \qquad \xi^{t-1} \left(\frac{t-1}{n-1} \cdot b\right)^{n-1} \cdot f \leq 1.$$

Substitute $t-1 = g(n-1)$. This transforms inequality (2) into

$$(\xi^g gb)^{n-1} \leq \frac{1}{f}.$$

Let g be such that $\xi^g \cdot g \cdot b \leq \min(1, 1/f)$. Take c so large that if $t > cn$ then $t-1 \geq g(n-1)$, because then

$$\xi^{t-1} \left(\frac{t-1}{n-1} b\right)^{n-1} f = (\xi^g gb)^{n-1} f \leq \xi^g gbf \leq 1,$$

for $n > 1$. □

Lemma 2.4 shows that from some moment the distribution of the time S needed to sort n 0's and 1's is bounded by the geometric distribution with parameter d. The time period of length cn is called the *principal time*, and, if $S > cn$, then $S - cn$ is the *delay time*. It follows from Lemma 2.4 that the delay time can be bounded by a geometric distribution with parameter d. Let n denote the number of processors. Then the sum of all the principal times over a branch is $c(n/2 + n/2^2 + \cdots + 1) = cn$, and the overall time for a branch is cn plus the sum of the delay times of all the nodes.

2.2. Tree process. In this section, we consider the maximum of the delay times over the branches of the tree of queues. The problem can be considered in the following generic form. Suppose that there is a full binary tree of height k, and a *node process* is associated with each node, which is a sequence of Bernoulli trials, each trial with probability δ of success. The whole tree defines the *tree process* as follows: It starts with the initialization of the node process of the root, and, generally if a node process terminates, then immediately the node processes at its children are started (if there are any). The tree process terminates when all the node processes have terminated. Let $G = G_n$ be the random variable equal to the time that the tree process must terminate, where $n = 2^k$. The following notation is used in the following:

$$P_k(t) = Pr(G \leq t),$$

$$R_k(t) = 1 - P_k(t) = Pr(G > t).$$

Throughout the paper, the notation $\lg x$ means $\log_2 x$.

LEMMA 2.5. *The following inequality holds*:

$$R_k(t) \leq (1-\delta)^{t-1} + 2\delta \sum_{i=1}^{t-1} (1-\delta)^{i-1} R_{k-1}(t-i),$$

for $k > 1$.

Proof. It follows from the definition of the tree process that numbers $P_k(t)$ satisfy the following recursive equality:

$$P_k(t) = \sum_{i=1}^{t-1} \delta(1-\delta)^{i-1} \cdot P_{k-1}^2(t-i).$$

Substitute $1 - R_i(x)$ for $P_i(x)$ to obtain

$$\begin{aligned} 1 - R_k(t) &= \sum_{i=1}^{t-1} \delta(1-\delta)^{i-1}[1 - R_{k-1}(t-1)]^2 \\ &= \sum_{i=1}^{t-1} \delta(1-\delta)^{i-1} - 2\delta \sum_{i=1}^{t-1} (1-\delta)^{i-1} R_{k-1}(t-i) + \Delta, \end{aligned}$$

where $\Delta = \sum_{i=1}^{t-1} \delta(1-\delta)^{i-1} R_{k-1}^2(t-i) \geq 0$. Rearrange the terms to obtain

$$R_k(t) = 1 - \sum_{i=1}^{t-1} \delta(1-\delta)^{i-1} + 2\delta \sum_{i=1}^{t-1} (1-\delta)^{i-1} R_{k-1}(t-i) - \Delta.$$

Use the equality

$$1 - \sum_{i=1}^{t-1} \delta(1-\delta)^{i-1} = (1-\delta)^{t-1}$$

and discard $-\Delta$ to obtain the inequality. □

LEMMA 2.6. *There are two constants* $x > 0$ *and* $y > 1$, *depending on* δ, *such that the following inequality holds*:

$$R_k(t) \leq 2^{kx} \cdot y^{-t}. \tag{3}$$

Proof. The proof is by induction on k, and simultaneously we stipulate the conditions that x and y must satisfy for the proof to be correct.

If $k = 1$, then $R_1(t) = (1-\delta)^t$. Inequality (3) holds if y satisfies the inequality $y(1-\delta) < 1$. Take y such that $1 < y < 1/(1-\delta)$.

Let $k > 1$. By Lemma 2.4 and the inductive assumption,

$$R_k(t) \leq (1-\delta)^{t-1} + 2\delta \sum_{i=1}^{t-1} (1-\delta)^{i-1} \cdot 2^{(k-1)x} \cdot y^{-t+i}$$
$$= (1-\delta)^{t-1} + 2^{kx} \cdot y^{-t} \cdot 2^{1-x} \cdot \delta \cdot y$$
$$\cdot \frac{1-[(1-\delta)y]^{t-1}}{1-(1-\delta)y} \leq (1-\delta)^{t-1} + 2^{kx} \cdot y^{-t} \cdot z,$$

where

$$z = \frac{2^{1-x}\delta y}{1-(1-\delta)y}.$$

We need the following inequality to be true:

$$(1-\delta)^{t-1} + 2^{kx} \cdot y^{-t} \cdot z \leq 2^{kx} \cdot y^{-t}.$$

This is equivalent to

$$((1-\delta)y)^t \cdot 2^{-kx} \leq (1-\delta)(1-z).$$

The left-hand side of the last inequality is largest for $t = 1$. For this value, the following is equivalent:

$$y2^{-kx} \leq 1 - z. \qquad (4)$$

Note that, for sufficiently large x, the inequality $1 - z > 0$ holds; namely,

$$1 < z \text{ iff } 2^{1-x} \cdot \frac{\delta y}{1-(1-\delta)y} < 1 \text{ iff } x > \lg\left(\frac{2\delta y}{1-(1-\delta)y}\right),$$

since $y(1-\delta) < 1$. Then inequality (4) is equivalent to

$$2^{kx} \geq \frac{y}{1-z}.$$

Take x sufficiently large to satisfy this inequality. □

COROLLARY 2.7. *If x and y are the same as in Lemma* 2.6, *then*

$$Pr(G_n > (x+t)\log_y n) \leq n^{-t}.$$

Proof. Denote $s = (x+t)\log_y n$. Then, by Lemma 2.6,

$$Pr(G_n > s) = R_k(s) \leq 2^{kx}y^{-s} = 2^{x \lg n} \cdot y^{-(x+t)\log_y n} = n^x n^{-x} n^{-t} = n^{-t}. \qquad \square$$

2.3. Back to the line of processors. The analysis of the tree of processes concerns the delay time G_n, the total time bounded by $cn + G_n$, where cn is the principal time as defined in §2.1.

THEOREM 2.8. *Let T denote the time needed to sort a line of n processors with delaying links. There are constants $r > 1$ and $D > 0$ such that the following is true*:

$$Pr(T > Dn + t) < r^{-t}.$$

Proof. Let numbers x and y be as in Lemma 2.6. Take for r any number satisfying $1 < r < y$. Number D will be specified later, but it will be greater than c, where cn is the principal time. By the definition of the principal time, the following inequality holds:

$$Pr(T > Dn + t) < Pr(G_n > (D - c)n + t).$$

Denote $s = (D - c)n + t$. Then, by Corollary 2.7,

$$Pr(G_n > s) < n^{-(s/\log_y n - x)} = 2^{-s\cdot\lg n/\log_y n + x\lg n} = 2^{-s\lg y + x\lg n} = y^{-s}\cdot n^x.$$

The inequality $y^{-s}n^x \le r^{-t}$ holds if and only if $y^{-(D-c)n}n^x \le \left(\frac{y}{r}\right)^t$. Take

$$D = c + \frac{x}{\lg y} \ge c + \frac{x\lg n}{n\lg y}$$

because then $y^{-(D-c)n}\cdot n^x \le 1 \le (y/r)^t$. □

The following terminology, extending that from §2.1, will be used in the following to refer to bounds like those in Theorem 2.8. Suppose that X_n is a random variable equal to the length of a time period depending on n and also that an *exponential inequality*

$$Pr(X_n > f(n) + t) < r^{-t}$$

holds. Then $f(n)$ is called the *principal time*, t is the *delay time*, and r is the *exponential factor*.

3. The two-dimensional case. In this section, we discuss sorting on a square mesh of processors interconnected by delaying links. If we take an algorithm that has the operations of sorting of rows or columns as its basic steps, then we expect it to sort in time $O(n)$ because the analysis from the previous section guarantees that each single row or column is sorted quickly. There are many such algorithms known in the literature, among them those developed by Lang et al. [8], Ma, Sen, and Scherson [12], or Schnorr and Shamir [14]. To be specific, we consider a simple algorithm of Leighton described in [11]. It sorts in the snake-like order, that is, the rows are ordered to the right and to the left, alternatingly. The *global snake* is an indexing scheme of processors corresponding to the snake-like ordering. The following is the algorithm that we later analyse.

Sorting Algorithm

1. Recursively sort each quadrant in the snake-like order.
2. Sort the rows alternatingly to the right and to the left.
3. Sort the columns.
4. Perform $4n$ steps of the odd-even transposition sort along the global snake.

Steps 2–4 are referred to as merging. This algorithm sorts in $O(n)$ time if all the links are fully operational (cf. [11]). In the situation when the links are delaying, a next step of the algorithm cannot start before the previous one has been completed. To control this, the following mechanism can be used. Consider, for instance, the step of sorting all the rows in parallel. Designate the leftmost processors as *control processors* and the top leftmost one as *master processor.* In each row, the rightmost processor sends to the left not only the key but also an additional control packet. This packet travels to the left and verifies if the row is sorted by comparing its value with the encountered keys. If the ordering has henceforth been correct, then the packet remembers the current key and is moved to the left; otherwise, it gets destroyed. When a control packet eventually reaches the control processor, then it testifies that the whole row has been sorted. The master processor learns about all the rows being sorted by receiving a packet originated by the bottom control processor and then forwarded and confirmed by the other control processors. The master processor then sends a signal to all the processors to start the next step.

We prove that this algorithm sorts the mesh in time $O(n)$ with large probability. In the next sections, the consecutive steps of the algorithm are investigated.

3.1. Merging. Suppose that there are four $k/2 \times k/2$ submeshes, where $k \leq n$, and that they are merged by steps 2–4 of the algorithm. The first phase is sorting k rows or columns. Let U_j be the delay time to sort the jth column, where $1 \leq j \leq k$. We need to estimate the distribution of

$$U = \max\{U_j | 1 \leq j \leq k\}.$$

It follows from Theorem 2.8 that

$$Pr(U_j < t) > 1 - r^{-t}$$

for some $r > 1$. Hence

$$Pr(U < t) > (1 - r^{-t})^n > 1 - nr^{-t}.$$

We would rather have an inequality of the form

$$Pr(U < t) > 1 - s^{-t}$$

for some $s > 1$ and sufficiently large t. Take $s = (r+1)/2$. Then

$$Pr(U < t) > 1 - s^{-t} \quad \text{iff} \quad t \cdot \lg\left(\frac{2r}{r+1}\right) > \lg n.$$

Therefore there is a constant c such that, if $t > c \lg n$, then

$$Pr(U < t) > 1 - s^{-t},$$

for $s > 1$. The time period $c \lg n$ is a contribution to the overall principal time.

The next phase of merging is the $4k$ steps of the odd-even transposition sort along the global snake. Divide the snake into intervals of length $4k$ and assume that the algorithm is performed only within the intervals. Next, shift the intervals by $2k$, execute the algorithm, and finally execute the algorithm once more on the original partition. Similarly as for the phase of sorting rows and columns in parallel, we obtain an exponential bound on the delay time with the principal time being $O(\lg n)$. Also, the coordination of steps by the master processor can be handled in the same way.

3.2. Recursion. The algorithm can be implemented bottom-up, that is, the submeshes are sorted and next merged into larger and larger ones. There are $\lg n$ levels. We estimate the maximum time of merging on a level. The obtained bounds are again exponential with only polylogarithmic principal time.

We have proved that the time T to merge four $k \times k$ submeshes satisfies the inequality

$$Pr(X > ck + t) < r^{-t}$$

for some $c > 0$ and $r > 1$. Let r_i denote the exponential factor on the level i. Define $r_0 = r$ and $r_i = r - i(r/2\lg n)$, for $i = 1, 2, \ldots, \lg n$. If there are l merges on level i and U_j denotes the delay time of the jth merge, then the overall delay time on this level $U = \max\{U_j | 1 \leq j \leq l\}$ is estimated as follows:

$$Pr(U < t) > (1 - r_i^{-t})^l > 1 - n^2 r_i^{-t}.$$

This can be extended to $1 - n^2 \cdot r_i^{-t} > 1 - r_{i+1}^{-t}$ for sufficiently large t; namely,

$$1 - n^2 \cdot r_i^{-t} > 1 - r_{i+1}^{-t} \quad \text{iff} \quad t > \frac{2 \ln n}{\ln\left(\frac{r_i}{r_{i+1}}\right)}.$$

However,

$$\frac{r_i}{r_{i+1}} = \frac{r - \dfrac{ir}{2\lg n}}{r - \dfrac{(i+1)r}{2\lg n}} = 1 + \frac{1}{2\lg n - i - 1} < 1 + q\frac{1}{\lg n}.$$

Since $\ln(1 + x) \leq x$, for $0 < x < 1$, we obtain $t > 2 \ln 2 \lg^2 n$. Therefore there is a polylogarithmic contribution to the principal time at each level.

The delay times are summed over the levels: If X_i is the delay on level i, then the total delay is $T = \sum_{i=1}^{\lg n} X_i$. Since $Pr(X_i > t) < (r/2)^{-t}$, for $1 \leq i \leq \lg n$, we can assume that all the distributions of X_i's are bounded by some geometric distribution with parameter p, $0 < p < 1$, specifically,

$$Pr(X_i = t) \leq p(1 - p)^t.$$

Then

$$Pr(T = t) \leq \begin{pmatrix} \lg n + t - 1 \\ t \end{pmatrix} p^{\lg n} \cdot (1 - p)^t.$$

The general bound

$$\begin{pmatrix} n \\ k \end{pmatrix} \leq \frac{n^n}{k^k (n - k)^{n-k}}$$

yields

$$\begin{pmatrix} x + y \\ x \end{pmatrix} \leq \left(1 + \frac{x}{y}\right)^y \cdot \left(1 + \frac{y}{x}\right)^x.$$

From this, we obtain

$$\begin{pmatrix} t + \lg n - 1 \\ t \end{pmatrix} \leq \left(1 + \frac{\lg n - 1}{t}\right)^t \cdot \left(1 + \frac{t}{\lg n - 1}\right)^{\lg n - 1}.$$

Use $(1+x/t)^t \leq e^x$ to transform this into

$$\begin{pmatrix} t+\lg n-1 \\ t \end{pmatrix} \leq e^{\lg n-1} \cdot \left(1+\frac{t}{\lg n-1}\right)^{\lg n-1} .$$

Substitute it in the bound on $Pr(T=t)$ to obtain

$$Pr(T=t) \leq \left[ep\left(1+\frac{t}{\lg n-1}\right)\right]^{\lg n}(1-p)^t.$$

We would like it to be bounded by q^t for some $0<q<1$ and sufficiently large t. Take any $1-p<q<1$, such as $q=1-p/2$. Then the inequality

$$\left[ep\left(1+\frac{t}{\lg n-1}\right)\right]^{\lg n}(1-p)^t<q^t$$

holds only if the following holds:

$$t>-\ln\left(\frac{1-p}{q}\right)\lg n \ln(ept),$$

for $t \geq 2$ and $n \geq 8$, because then $1+t/(\lg n-1) \leq t$. Take $t>\ln^2 n$. Then

$$\frac{t}{\ln(ept)}>\frac{\lg^2 n}{\ln(ep\lg^2 n)}>-\ln\left(\frac{1-p}{q}\right)\lg n.$$

Hence $Pr(T=t) \leq q^t$ for $t>\lg^2 n$. From this, it follows that

$$Pr(T>\lg^2 n+t) \leq \sum_{i=\lg^2 n+t}^{\infty} q^i<q^t,$$

this being valid for sufficiently large n. This formula contributes only $O(\lg^2 n)$ to the principal time.

THEOREM 3.1. *There are constants $c>0$ and $r>1$ such that an $n \times n$ mesh with delaying links can be sorted in time $c \cdot n+t$ with the probability at least $1-r^{-t}$.*

Proof. The time to sort the mesh is a sum of some finite fixed set of random variables, each depending on n and satisfying an exponential inequality with at most linear principal time. □

4. Conclusions and open problems. We considered the problem of sorting on a mesh-connected processor array in which every attempt of communication between two neighboring processors fails with a constant probability and all such failures are independent of each other. It was proved that an $n \times n$ mesh can be sorted in the expected time $O(n)$. Although we considered one specific sorting algorithm, it follows from the presented analysis that many other known algorithms can be readily adapted to sort in time $O(n)$ on the model under consideration. This holds, for instance, for the algorithms developed in [8], [10], [12]–[15].

The results presented in this paper can be extended to include sorting on multidimensional meshes (cf. [7], [9], [12], [15]). It would be also interesting to consider sorting and related problems for other architectures with delaying links, such as hypercubes or butterflies.

REFERENCES

[1] P. BERMAN AND J. SIMON, *Investigations of fault-tolerant networks of computers*, in Proc. 20th Annual ACM Sympos. on Theory of Computing, 1988, pp. 66–77.

[2] B. S. CHLEBUS, K. DIKS, AND A. PELC, *Fast gossiping with short unreliable messages*, Discrete Applied Math., to appear.

[3] D. GROSS AND C. M. HARRIS, *Fundamentals of Queueing Theory*, John Wiley, New York, 1985.

[4] J. HASTAD, T. LEIGHTON, AND M. NEWMAN, *Fast computation using faulty hypercubes*, in Proc. 21st Annual ACM Sympos. on Theory of Computing, 1989, pp. 251–263.

[5] C. KAKLAMANIS, A. R. KARLIN, F. T. LEIGHTON, V. MILENKOVIC, P. RAGHAVAN, S. RAO, C. THOMBORSON, AND A. TSANTILAS, *Asymptotically tight bounds for computing with faulty arrays of processors*, in Proc. 31st Annual IEEE Sympos. on Foundations of Computer Science, 1990, pp. 285–296.

[6] J. G. KEMENY, J. L. SNELL, AND A. W. KNAPP, *Denumerable Markov Chains*, Springer-Verlag, New York, 1976.

[7] M. KUNDE, *Optimal sorting on multi-dimensionally mesh-connected computers*, in Proc. STACS'87, Lecture Notes in Computer Science, 247 (1987), pp. 408–419.

[8] H. W. LANG, M. SCHIMMLER, H. SCHMECK, AND H. SCHRODER, *Systolic sorting on a mesh-connected network*, IEEE Trans. Comput., C-34 (1985), pp. 652–658.

[9] F. T. LEIGHTON, *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*, Morgan Kaufmann, 1992.

[10] ———, *Tight bounds on the complexity of parallel sorting*, IEEE Trans. Comput., C-34 (1985), pp. 344–354.

[11] F. T. LEIGHTON, C. E. LEISERSON, AND D. KRAVETS, *Advanced Parallel and* VLSI *Computation: Lecture Notes for* 18.435/6.848, MIT/LCS/RSS 8, 1990.

[12] Y. MA, S. SEN, AND I. D. SCHERSON, *The distance bound for sorting on mesh-connected processor array is tight*, in Proc. 27th Annual IEEE Sympos. on Foundations of Computer Science, 1986, pp. 255–263.

[13] D. NASSIMI AND S. SAHNI, *Bitonic sort on a mesh-connected parallel computer*, IEEE Trans. Comput., C-28 (1979), pp. 2–7.

[14] C. P. SCHNORR AND A. SHAMIR, *An optimal sorting algorithm for mesh connected computers*, in Proc. 18th Annual ACM Sympos. on Theory of Computing, 1986, pp. 255–263.

[15] C. D. THOMPSON AND H. T. KUNG, *Sorting on a mesh-connected parallel computer*, Comm. Assoc. Comput. Mach., 20 (1977), pp. 263–271.

SIAM J. DISCRETE MATH
Vol. 7, No. 1, pp. 133–140, February 1994

LABELING CHORDAL GRAPHS: DISTANCE TWO CONDITION*

DENISE SAKAI[†]

Abstract. An $L(2,1)$-*labeling* of a graph G is an assignment of nonnegative integers to the vertices of G such that adjacent vertices get numbers at least two apart, and vertices at distance two get distinct numbers. The $L(2,1)$-*labeling number* of G, $\lambda(G)$, is the minimum range of labels over all such labelings. It is shown that, for chordal graphs G with maximum degree $\triangle(G)$, $\lambda(G) \leq (\triangle(G)+3)^2/4$; in particular, if G is a unit interval graph with chromatic number $\chi(G)$, $\lambda(G) \leq 2\chi(G)$, which is a better bound. As a consequence, it is shown that the conjecture $\lambda(G) \leq \triangle^2(G)$ by Griggs and Yeh [*SIAM J. Discrete Math.*, 5 (1992), pp. 586–595] is true for chordal graphs.

Key words. labeling, T-coloring, chordal graphs, unit interval graphs

AMS subject classification. 05C78

1. Introduction. In the *channel assignment problem*, we wish to assign a channel (nonnegative integer) to each TV or radio transmitter such that interfering transmitters get channels whose separation is not in a set of disallowed separations. This problem was first formulated as a graph coloring problem by Hale [4], who introduced the notion of *T-coloring* of a graph. Here, the transmitters are represented by the vertices of a graph, called the *interference graph*, the edges represent interference, the channels are the colors assigned to the vertices, and T is the set of disallowed separations. For recent surveys about T-colorings and the channel assignment problem, see Roberts [6], [7] and Tesman [8].

In 1988, Roberts (in a private communication to Griggs) proposed a variation of the channel assignment problem, where "close" transmitters must receive different channels and "very close" transmitters must receive channels at least two apart. Griggs and Yeh [3] and Yeh [9] consider a more general problem. Given a real number $d > 0$, an $L_d(2,1)$-*labeling* of a graph G is an assignment f of nonnegative real numbers to the vertices of G such that, if x and y are two adjacent vertices, then $|f(x) - f(y)| \geq 2d$, and, if the distance between x and y is two, then $|f(x) - f(y)| \geq d$.

Griggs and Yeh [3] concentrate their attention on the $L_d(2,1)$-*labeling number* of G, denoted by $\lambda(G,d)$, the smallest number m such that G has an $L_d(2,1)$-labeling with no label greater than m. If f is an $L_d(2,1)$-labeling of G, we say that $f \in L_d(2,1)(G)$, and then

$$\lambda(G,d) = \min_{f \in L_d(2,1)(G)} ||f(G)||,$$

where $||f(G)|| = \max_{v\in V(G)} f(v)$. $\lambda(G,d)$ is sometimes called the *span* and is the minimum separation between the largest and the smallest channels used. Griggs and Yeh characterize $\lambda(G,d)$ in terms of $\lambda(G,1)$ and find that for $\lambda(G,1)$ it suffices to consider labelings using only nonnegative integers. So, in what follows, we focus on $L_1(2,1)$-labelings using only nonnegative integers. For simplicity, we denote $L_1(2,1)(G)$ by $L(2,1)(G)$, and $\lambda(G,1)$ by $\lambda(G)$. Griggs and Yeh [3] find exact values for $\lambda(G)$ for particular graphs G.

*Received by the editors December 9, 1991; accepted for publication (in revised form) November 6, 1992. This work was supported by Air Force Office of Scientific Research grants AFOSR-89-0512 and AFOSR-90-0008, by FAPESP—Fundação de Amparo à Pesquisa do Estado de São Paulo—Brazil, and by the University of São Paulo–Brazil.

[†]Department of Mathematics and Computer Science, Montclair State College, Upper Montclair, New Jersey 07043 (sakai@tempest.montclair.edu).

PROPOSITION 1.1 (Griggs and Yeh [3]). (i) *Let P_n be a path on $n \geq 2$ vertices. Then*

$$\lambda(P_n) = \begin{cases} 2, & \text{if } n = 2, \\ 3, & \text{if } n = 3, 4, \\ 4, & \text{if } n \geq 5. \end{cases}$$

(ii) *Let C_n be a circuit on $n \geq 3$ vertices. Then $\lambda(C_n) = 4$.*

For some other families of graphs, they find interesting bounds for $\lambda(G)$. To mention a few, $\lambda(T) \leq \triangle(T) + 2$, if T is a tree with maximum degree $\triangle(T) \geq 1$; $\lambda(Q_n) \leq 2n+1$, if Q_n is an *n-cube*, i.e, a graph with vertices given by the n-tuples with 0,1 coordinates, and with two vertices adjacent if they differ in only one coordinate.

They also propose the following interesting conjecture.

CONJECTURE 1.2 (Griggs and Yeh [3]). *For any graph G with maximum degree $\triangle(G) \geq 2$, $\lambda(G) \leq \triangle^2(G)$.*

Many classes of graphs satisfy this conjecture, for instance, the class of graphs with diameter 2 considered by Griggs and Yeh [3]. In §2 we show that one more important class of graphs satisfies Conjecture 1.2, namely, the *chordal* graphs, i.e., graphs that do not contain induced circuits with more than three vertices. A better bound for $\lambda(G)$ is obtained if we restrict our attention to a particular subfamily of the chordal graphs, namely, the unit interval graphs.

The class of unit interval graphs and its generalization, the class of R-unit sphere graphs, are of particular interest in the channel assignment problem. A graph is *R-unit sphere* if we can assign a closed sphere of unit diameter in R-space to each vertex so that edges correspond to pairs of spheres that overlap. When transmitters are located in R-space, for R=1, 2, or 3, interference sometimes takes place if and only if two transmitters are within m miles. In this case, the interference graph is an R-unit sphere graph. Unfortunately, no useful characterization of R-unit sphere graphs is known except when R = 1. The 1-unit sphere graphs are sometimes called *unit interval graphs* since a closed sphere of unit diameter in 1-space is simply an interval of unit length. So the unit interval graphs can be interpreted as interference graphs of transmitters located in a linear corridor, where interference corresponds to being within m miles. In §2, in addition to finding an upper bound for $\lambda(G)$ when G is unit interval, we also characterize unit interval graphs with a certain number of vertices, for which this bound is tight.

2. L(2,1)-labelings of unit interval graphs. We begin this section by presenting an upper bound for $\lambda(G)$, for G a chordal graph. First, let us recall an important property of chordal graphs due to Dirac. A *simplicial vertex* of a graph is a vertex such that its neighbors induce a clique in the graph.

THEOREM 2.1 (Dirac [1]). *If G is a chordal graph, then it has a simplicial vertex.*

In particular, if G is a chordal graph and v is a simplicial vertex, then the graph $G-v$, obtained from G by removing vertex v and all the edges incident to v, is also chordal.

THEOREM 2.2. *Let G be a chordal graph with maximum degree $\triangle(G)$. Then $\lambda(G) \leq (\triangle(G) + 3)^2/4$.*

Proof. The proof is by induction on the number of vertices of G, $|V(G)|$. If $|V(G)| = 1$, then $\lambda(G) = 0 = \triangle(G)$, and obviously $\lambda(G) \leq (\triangle(G)+3)^2/4$. Suppose that $|V(G)| > 1$ and that G is chordal. By Theorem 2.1, since G is chordal, it has a simplicial vertex, say v, and $G - v$ is also chordal. We can apply induction to conclude that

$$\lambda(G - v) \leq \frac{(\triangle(G - v) + 3)^2}{4} \leq \frac{(\triangle(G) + 3)^2}{4}.$$

Let $f \in L(2,1)(G - v)$ with $||f(G - v)|| = \lambda(G - v)$ and denote by k the degree of v in G. Then v is distance one away from k vertices and distance two away from at most $k(\triangle(G) - k)$ vertices (recall that v and its neighbors form a clique in G). Therefore there are at most $3k + k(\triangle(G) - k)$ numbers used by f to be avoided by v. However, the function $g(k) = 3k + k(\triangle(G) - k)$ has its maximum at $k = (\triangle(G) + 3)/2$ and $g((\triangle(G) + 3)/2) = (\triangle(G) + 3)^2/4$; i.e., there are at most $\lfloor(\triangle(G) + 3)^2/4\rfloor$ numbers used by f to be avoided by v. Hence there is still at least one number in $\{0, 1, ..., \lfloor(\triangle(G) + 3)^2/4\rfloor\}$ to be assigned to v to extend f into a labeling in $L(2,1)(G)$. □

Unfortunately, we do not have examples where this upper bound is sharp.

As a corollary, we have that chordal graphs with maximum degree greater than 1 satisfy Conjecture 1.2.

COROLLARY 2.3. *Let G be a chordal graph with maximum degree* $\triangle(G) \geq 2$. *Then* $\lambda(G) \leq \triangle^2(G)$.

Proof. Let G be a chordal graph with $\triangle(G) \geq 2$. If $\triangle(G) = 2$, then G is a disjoint union of paths and triangles. So, by Proposition 1.1, $\lambda(G) \leq 4 = \triangle^2(G)$. If $\triangle(G) \geq 3$, then $(\triangle(G) + 3)^2/4 \leq \triangle^2(G)$, and therefore, by Theorem 2.2, $\lambda(G) \leq \triangle^2(G)$. □

In the following, we use the notation $n(G)$, $\chi(G)$, and $\omega(G)$ to denote the number of vertices, the chromatic number, and the size of the maximum clique of the graph G, respectively. Where there is no possibility of confusion, we write $n = n(G)$, $\chi = \chi(G)$, and $\omega = \omega(G)$.

We concentrate on a particular family of chordal graphs, the unit interval graphs, and, in the next theorem, we improve the bound for $\lambda(G)$ given by Theorem 2.2. Roberts [5] showed that a graph G is unit interval if and only if it has a *compatible vertex ordering*, i.e., an ordering $v_1, v_2, ..., v_n$ of vertices in G so that, if $i < j < k$ and $\{v_i, v_k\}$ is an edge, then $\{v_i, v_j\}, \{v_j, v_k\}$ are edges in G. With this characterization, it is not difficult to see that unit interval graphs are chordal. It is worth noting that, if G is a unit interval, then $\chi(G) = \omega(G)$ can be seen easily by greedily (proper) coloring the vertices in the compatible ordering.

THEOREM 2.4. *Suppose that G is a unit interval graph. Then*

$$2\chi(G) - 2 \leq \lambda(G) \leq 2\chi(G).$$

Proof. Let G be a unit interval graph and suppose that $v_1, v_2, ..., v_n$ is a compatible vertex ordering. Start labeling the vertices from the beginning of the compatible vertex ordering in order, with the numbers in the following sequence:

$$2\chi - 2, 2\chi - 4, ..., 2, 0, 2\chi - 1, 2\chi - 3, ..., 3, 1, 2\chi.$$

If the sequence of numbers finishes and there are vertices still unlabeled (i.e., if $n(G) > 2\chi(G) + 1$), continue labeling, repeating the same sequence until no more vertices remain to be labeled.

Let us show that the above labeling is in $L(2,1)(G)$. Only two possible problems could occur. We now investigate them.

Problem 1. A vertex u labeled $2i$ with $i \in \{0, 1, ..., \chi(G)\}$ is adjacent to a vertex v labeled $2i \pm 1$: We show that this problem cannot occur. If $u = v_q$, then the first vertex following u labeled $2i \pm 1$, if there is such a vertex, is $v_{q+\chi}$. So, if v_q and $v_{q+\chi}$ are adjacent, then the vertices $v_q, v_{q+1}, ..., v_{q+\chi}$ form a clique of size $\chi(G) + 1$, which is an impossibility, since $\chi(G) \geq \omega(G)$ for any G. Thus v_q and $v_{q+\chi}$ are not adjacent; then v cannot follow u and be adjacent to it, since otherwise v will follow $v_{q+\chi}$ and force the adjacency of v_q and $v_{q+\chi}$. On the other hand, the vertex labeled $2i \pm 1$ preceding v_q that is closest to v_q, if there is any, is $v_{q-\chi}$. If v_q and $v_{q-\chi}$ are adjacent, however, then $v_{q-\chi}, v_{q-\chi+1}, ..., v_q$

form a clique of size $\chi(G)+1$, a contradiction. Then v cannot precede u and be adjacent to it, since otherwise v will precede $v_{q-\chi}$ and force the adjacency of v_q and $v_{q-\chi}$.

Problem 2. Two vertices at distance two get the same label: Let u and v be two vertices at distance two and, without loss of generality, assume that $u = v_q$ and that v follows u in the compatible vertex ordering. If v_q has label $i \in \{0,1,...,2\chi(G)\}$, then the first vertex following v_q with label i, if there is any, is $v_{q+2\chi+1}$. So $v = v_p$ for some $p \geq q+2\chi(G)+1$. Since u and v are at distance two, then there is a vertex v_r such that v_q and v_p are adjacent to v_r. Now $q < r < p$, for otherwise, say, without loss of generality, that $r > p$. Then v_r adjacent to v_q implies v_p adjacent to v_q, contrary to the assumption that v_p and v_q are at distance two. Therefore either $v_q, v_{q+1}, ..., v_r$ or $v_r, v_{r+1}, ..., v_p$ is a clique of size greater than $\chi(G)$ (because $p \geq q + 2\chi(G) + 1$), a contradiction. Hence Problem 2 cannot occur either.

Note that the labeling constructed above has span $2\chi(G)$; hence $\lambda(G) \leq 2\chi(G)$. Obviously, $\lambda(G) \geq 2\chi(G) - 2$ since G contains a clique of size $\chi(G)$. □

An immediate corollary of the proof of Theorem 2.4 is the following.

COROLLARY 2.5. *Suppose that G is a unit interval graph with* $n(G) < 2\chi(G)+1$. *Then* $2\chi(G) - 2 \leq \lambda(G) \leq 2\chi(G) - 1$.

Since, for any graph G, $\chi(G) \leq \Delta(G) + 1$, it follows from Theorem 2.4 that, if G is unit interval and $\Delta(G) \geq 0$, then

$$2\chi(G) - 2 \leq \lambda(G) \leq 2\chi(G) \leq 2(\Delta(G) + 1) \leq \frac{(\Delta(G) + 3)^2}{4},$$

improving the upper bound given by Theorem 2.2 for general chordal graphs when $\Delta(G) \neq 1$. Also, note that $\lambda(G) \geq 2\chi(G) - 2$ holds for all perfect graphs, not just unit interval graphs.

The lower bound for $\lambda(G)$ given in Theorem 2.4 is attained, for instance, when G is a complete graph. In Proposition 2.7, we show that the upper bound for $\lambda(G)$ is also tight.

LEMMA 2.6. *Let* $f \in L(2,1)(G)$ *with* $||f(G)|| = t$. *If v is a vertex of degree* $t-1$, *then* $f(v) \in \{0, t\}$.

Proof. The proof is straightforward, and we omit the details. □

An *r-path* is a graph with vertex set $v_1, v_2, ..., v_n$, $n > r$, and $v_i, v_{i+1}, ..., v_{i+r}$ is a clique, for $i = 1, 2, ..., n-r$. Clearly, an r-path is unit interval, has chromatic number $r+1$, and $v_1, v_2, ..., v_n$ is a compatible vertex ordering.

PROPOSITION 2.7. *Let G be an r-path with* $2r+3$ *vertices. Then* $\lambda(G) = 2(r+1) = 2\chi(G)$.

Proof. If $r = 1$, then G is a path on five vertices and, by Proposition 1.1, $\lambda(G) = 4 = 2\chi(G)$. Suppose that $r > 1$ and assume, by way of contradiction, that $\lambda(G) \leq 2\chi(G) - 1$. Let $f \in L(2,1)(G)$ with $||f(G)|| = 2\chi(G) - 1$.

By Lemma 2.6, since $v_\chi, v_{\chi+1}, v_{\chi+2}$ have degree $2r = 2\chi(G) - 2$, they must get labels in $\{0, 2\chi(G) - 1\}$. However, $r > 1$ forces $v_\chi, v_{\chi+1}, v_{\chi+2}$ to be a triangle, and it is not possible to label it with only two distinct numbers. Therefore $\lambda(G) = 2\chi(G)$. □

It seems to be difficult to characterize those unit interval graphs G with $\lambda(G) = 2\chi(G)$. By Corollary 2.5, we know that such graphs must have at least $2\chi(G)+1$ vertices. The next few results give a characterization of unit interval graphs G with $\lambda(G) = 2\chi(G)$ and $n(G) = 2\chi(G) + 1$.

LEMMA 2.8. *Let G be a unit interval graph with* $n(G) = 2\chi(G) + 1$ *vertices. If there is a compatible vertex ordering* $v_1, v_2, ..., v_n$ *of the vertices such that either*

(i) $\{v_1, v_\chi\} \notin E(G)$, *or*
(ii) $\{v_{\chi+2}, v_n\} \notin E(G)$, *or*
(iii) $\{v_2, v_{\chi+1}\}, \{v_3, v_{\chi+2}\}, ..., \{v_\chi, v_{2\chi-1}\} \notin E(G)$, *or*
(iv) $\{v_3, v_{\chi+2}\}, \{v_4, v_{\chi+3}\}, ..., \{v_{\chi+1}, v_{2\chi}\} \notin E(G)$,
then $\lambda(G) \leq 2\chi(G) - 1$.

Proof. If one of the above items is satisfied, then labeling $v_1, v_2, ..., v_n$ in order with the numbers in the sequence
(i) $1, 2\chi - 2, 2\chi - 4, ..., 2, 0, 2\chi - 1, 2\chi - 3, ...3, 1$,
(ii) $1, 3, ..., 2\chi - 3, 2\chi - 1, 0, 2, ..., 2\chi - 4, 2\chi - 2, 1$,
(iii) $2\chi - 1, 2\chi - 3, ..., 3, 1, 2\chi - 2, 2\chi - 4, ..., 2, 0, 2\chi - 1$, and
(iv) $2\chi - 1, 0, 2, ..., 2\chi - 4, 2\chi - 2, 1, 3, ..., 2\chi - 3, 2\chi - 1$,
respectively, we obtain a labeling in $L(2,1)(G)$. So $\lambda(G) \leq 2\chi(G) - 1$. □

Before presenting the next main result, Theorem 2.12, we must introduce further notation and some auxiliary results.

For each integer $k > 2$, let

$$K_i(k) = \{x_i(k), x_{i+1}(k), ..., x_{i+k-1}(k)\},$$

for $i = 1, 2, ..., k+1$, where

$$x_j(k) = 2(k - j), \qquad x_{k+j}(k) = 2(k - j) + 1,$$

for $j = 1, 2, ..., k$. Let H_k be the graph with vertex set

$$V(H_k) = \{x_1(k), x_2(k), ..., x_{2k}(k)\}$$

so that

$$\{x, y\} \in E(H_k) \Leftrightarrow x, y \in K_i(k),$$

for some $i = 1, 2, ..., k+1$. Note that H_k is an r-path for $r = k - 1$ with $\{K_i(k) : i = 1, 2, ..., k+1\}$ as the set of all its maximal cliques. It is clear that the following property is satisfied.

PROPERTY 2.9. *If* $i < j < l$ *are integers and* $x \in K_i(k) \cap K_l(k)$, *then* $x \in K_j(k)$.

Actually, we should mention that Property 2.9 gives us a characterization of interval graphs. A graph G is an *interval graph* if we can assign an interval to each vertex so that edges correspond to pairs of intervals that overlap.

THEOREM 2.10 (Fulkerson and Gross [2]). *A graph is interval if and only if there is a ranking* $K_1, K_2, ..., K_p$ *of all its maximal cliques, which satisfies Property* 2.9.

The following lemma provides all the possible sets of labels to be assigned to the cliques of size k in an arbitrary graph G, by a labeling $f \in L(2,1)(G)$ with $||f(G)|| \leq 2k - 1$. This family of sets proves to have an interval graph structure.

LEMMA 2.11. *Let* G *be a graph and let* k *be an integer greater than* 2. *Suppose that there is an* $f \in L(2,1)(G)$ *such that* $||f(G)|| \leq 2k - 1$. *If* K *is a clique of* G *with size* k, *then* $f(K) = K_i(k)$ *for some* $i = 1, 2, ..., k+1$, *where* $f(K) = \{f(v) : v \in K\}$.

Proof. We argue by induction on k. If $k = 3$, then

$$x_1(3) = 4, \quad x_2(3) = 2, \quad x_3(3) = 0, \quad x_4(3) = 5, \quad x_5(3) = 3, \quad x_6(3) = 1,$$

and we can check that

$$K_1(3) = \{4, 2, 0\}, \quad K_2(3) = \{2, 0, 5\}, \quad K_3(3) = \{0, 5, 3\}, \quad K_4(3) = \{5, 3, 1\},$$

are all the possible sets of labels that can be associated by f to a clique of size 3.

Suppose that $k > 3$. Let K be a clique of size k. Since $\lambda(K) = 2k-2$, either $2k-1$ or $2k-2$, but not both, is in $f(K)$. Let G' be the graph obtained from G by removing all the vertices labeled $2k-1$ or $2k-2$ and the edges incident to them. More formally, $G' = G - X$, where $X = \{x \in V(G) : f(x) = 2k-1 \text{ or } 2k-2\}$. Let f' be the restriction of f to the vertices of G', i.e., $f'(x) = f(x)$ for $x \in V(G')$. Note that $f' \in L(2,1)(G')$ and $||f'(G')|| \leq 2k-3 = 2(k-1)-1$. Let $v \in K$ with $f(v) \in \{2k-1, 2k-2\}$. Then $K' = K - v$ is a clique of size $k-1$ in G'. By the induction hypothesis $f'(K') = K_i(k-1)$ for some $i = 1, 2, ..., (k-1)+1$. However, $K_j(k-1) \cap \{2k-1, 2k-2\} = \emptyset$, for $j = 1, 2, ..., (k-1)+1$ and

$$K_1(k) = K_1(k-1) \cup \{2k-2\},$$

$$K_{i+1}(k) = K_i(k-1) \cup \{2k-1\}, \qquad i = 1, 2, ..., k.$$

Therefore

$$f(K) = f(K' \cup \{v\}) = f'(K') \cup \{f(v)\} = K_j(k)$$

for some $j = 1, 2, ..., k+1$. □

Finally, the next result gives a characterization of unit interval graphs G on $n(G) = 2\chi(G) + 1$ vertices with $\lambda(G) = 2\chi(G)$. It essentially shows that the converse in Lemma 2.8 also holds.

THEOREM 2.12. *Let G be a unit interval graph on $n(G) = 2\chi(G) + 1$ vertices and $\chi(G) > 2$. There is a compatible vertex ordering $v_1, v_2, ..., v_n$ such that either*

(i) $\{v_1, v_\chi\}, \{v_{\chi+2}, v_n\}, \{v_q, v_{q+\chi-1}\} \in E(G)$ *for some* $3 \leq q \leq \chi(G)$, *or*

(ii) $\{v_1, v_\chi\}, \{v_2, v_{\chi+1}\}, \{v_{\chi+1}, v_{n-1}\}, \{v_{\chi+2}, v_n\} \in E(G)$,

if and only if $\lambda(G) = 2\chi(G)$.

Proof. Suppose first that $\lambda(G) = 2\chi(G)$ under the hypothesis of the theorem and choose any compatible vertex ordering. Observe that neither one of the items in the statement of Lemma 2.8 can occur. So one of the above items must occur.

Conversely, suppose that (ii) holds and by way of contradiction that there is $f \in L(2,1)(G)$ such that $||f(G)|| \leq 2\chi(G) - 1$. Since $v_1, v_2, ..., v_n$ is a compatible vertex ordering and $\{v_2, v_{\chi+1}\}, \{v_{\chi+1}, v_{n-1}\} \in E(G)$, the sets

$$A = \{v_2, v_3, ..., v_{\chi+1}\}, \qquad B = \{v_{\chi+1}, v_{\chi+2}, ..., v_{n-1}\}$$

are two cliques with $A \cap B = \{v_{\chi+1}\}$. Since $v_{\chi+1}$ has degree $2\chi - 2$, $||f(G)|| = 2\chi(G) - 1$, and $v_{\chi+1}$ must get label 0 or $2\chi(G) - 1$ by Lemma 2.6. Suppose that $f(v_{\chi+1}) = 0$ (the case where $f(v_{\chi+1}) = 2\chi(G) - 1$ is symmetric). By Lemma 2.11, the only possible sets of labels that can be assigned to A or B are

$$K_1(\chi(G)) = \{0, 2, 4, ..., 2\chi(G) - 2\},$$

$$K_\chi(\chi(G)) = \{0, 3, 5, ..., 2\chi(G) - 1\},$$

since the vertices in $A - \{v_{\chi+1}\}$ are at distance two of vertices in $B - \{v_{\chi+1}\}$, so $f(A) \cap f(B) = \{0\}$. Suppose that $f(A) = K_1(\chi(G))$. Since $\{v_1, v_\chi\} \in E(G)$, $\{v_1, v_2, ..., v_\chi\}$ is also a clique and $f(\{v_2, v_3, ..., v_\chi\}) = \{2, 4, ..., 2\chi(G) - 2\}$, the only possible label for v_1 is 0. However, v_1 is at distance two from $v_{\chi+1}$ labeled 0. Therefore $f(A)$ cannot be $K_1(\chi(G))$. Similarly, $f(B)$ cannot be $K_1(\chi(G))$, and we have reached a contradiction. Hence $\lambda(G) = 2\chi(G)$.

On the other hand, suppose that (i) holds and, by way of contradiction, assume that there is $f \in L(2,1)(G)$ with $||f(G)|| \leq 2\chi(G) - 1$. Since $\{v_1, v_\chi\}$, $\{v_{\chi+2}, v_n\}$, $\{v_q, v_{q+\chi-1}\} \in E(G)$, for some $3 \leq q \leq \chi(G)$, and $v_1, v_2, ..., v_n$ is a compatible vertex ordering,

$$A = \{v_1, v_2, ..., v_\chi\}, \quad B = \{v_{\chi+2}, v_{\chi+3}, ..., v_n\}, \quad C = \{v_q, v_{q+1}, ..., v_{q+\chi-1}\}$$

are cliques of size $\chi(G)$.

By the previous lemma, there are $i, j, l \in \{1, 2, ..., \chi(G) + 1\}$ such that $f(A) = K_i(\chi(G))$, $f(B) = K_j(\chi(G))$, and $f(C) = K_l(\chi(G))$. Since $3 \leq q \leq \chi(G)$, $A \cap C$, $C \cap B$, $A - C$, $B - C$, and $C - (A \cup B)$ are nonempty sets, and $A \cap B \cap C = \emptyset$. Clearly, $i \neq l \neq j$. Suppose that $i < l$ (the case where $i > l$ is similar). If $j < l$, either $l \leq \chi(G)$ or $i, j \geq 2$, since $K_i(\chi(G)) \cap K_l(\chi(G)) \neq \emptyset$ and $K_j(\chi(G)) \cap K_l(\chi(G)) \neq \emptyset$. Then, however, $K_i(\chi(G)) \cap K_j(\chi(G)) \cap K_l(\chi(G)) \neq \emptyset$. This forces the existence of a vertex $v \in C \cap B$ and a vertex $w \in A$ (or, symmetrically, $v \in C \cap A$ and $w \in B$) such that $f(v) = f(w)$. Note, however, that v and w are distinct, since $A \cap B \cap C = \emptyset$, and they are at distance at most two, so they cannot get the same label, a contradiction. Therefore we must have $i < l < j$. Since $n(G) = 2\chi(G) + 1$ and we have $2\chi(G)$ labels in $\{0, 1, ..., 2\chi(G) - 1\}$, there must exist a label y and two vertices v and w with this label. However, if $v \in C$, then the distance in G between v and w is at most two, so $f(v) \neq f(w)$. So $v \notin C$. Similarly, $w \notin C$. Then, without loss of generality, $v \in A$ and $w \in B$. Then $y \in K_i(\chi(G)) \cap K_j(\chi(G))$, and, by Property 2.9, since $i < l < j$, $y \in K_l(\chi(G))$. However, any vertex in C with label y is distinct from v and w, and it is at distance at most two from v and w, a contradiction. So $\lambda(G) = 2\chi(G)$. □

3. Further research. The ultimate objective of our work is to understand Griggs and Yeh's Conjecture 1.2, mentioned in the Introduction, by restricting ourselves to the class of chordal graphs and, in particular, to the class of unit interval graphs. This seems to be a reasonable way to approach a problem of this surprising difficulty, especially considering the fact that positive results for very simple classes of graphs are not easy to obtain. A natural class of graphs to be examined next would be the class of interval graphs.

Let us close by addressing a few more questions left unsolved in this paper.

(i) Is the upper bound of Theorem 2.2 sharp?

(ii) Generalize Theorem 2.12 for unit interval graphs G with more than $2\chi(G) + 1$ vertices.

(iii) Characterize unit interval graphs G with $\lambda(G) = 2\chi(G) - 2$ and with $\lambda(G) = 2\chi(G) - 1$ (from Theorem 2.4, we know that $2\chi(G) - 2 \leq \lambda(G) \leq 2\chi(G)$).

Acknowledgments. The author gratefully thanks Fred Roberts, Jerrold Griggs, and the referees for their helpful comments and suggestions. The author also thanks the Rutgers Center for Operations Research, Rutgers University, New Jersey.

REFERENCES

[1] G. A. Dirac, *On rigid circuit graphs*, Abh. Math. Sem. Univ. Hamburg, 25 (1961), pp. 71–76.

[2] D. R. Fulkerson and O. A. Gross, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835-855.

[3] J. R. Griggs and R. K. Yeh, *Labeling graphs with a condition at distance two*, SIAM J. Discrete Math., 5 (1992), pp. 586–595.

[4] W. K. Hale, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.

[5] F. S. Roberts, *On the compatibility between a graph and a simple order*, J. Combin. Theory, 11 (1971), pp. 28–38.

[6] ———, *From garbage to rainbows: Generalizations of graph coloring and their applications*, in Proc. 6th Internat. Conference on the Theory and Applications of Graphs, Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, eds., John Wiley, New York, 1990.

[7] ———, *T-Colorings of graphs: Recent results and open problems*, Discrete Math., 93 (1991), pp. 229–245.

[8] B. A. Tesman, *T-Colorings, List T-Colorings and Set T-Colorings*, Ph.D. thesis, Dept. of Math., Rutgers University, New Brunswick, NJ, 1989.

[9] R. K. Yeh, *Labeling Graphs with a Condition at Distance Two*, Ph.D. thesis, Dept. of Math., University of South Carolina, Spartanburg, SC, 1990.

SIAM J. DISCRETE MATH.
Vol. 7, No. 1, pp. 141–151, February 1994

POLYHEDRAL CHARACTERIZATION OF THE ECONOMIC LOT-SIZING PROBLEM WITH START-UP COSTS*

C. P. M. VAN HOESEL†, A. P. M. WAGELMANS‡, AND L. A. WOLSEY§

Abstract. A class of strong valid inequalities is described for the single-item uncapacitated economic lot-sizing problem with start-up costs. It is shown that these inequalities yield a complete polyhedral characterization of the problem. The corresponding separation problem is formulated as a shortest path problem. Finally, a reformulation as a plant location problem is shown to imply the class of strong valid inequalities, which shows that this reformulation is tight, also.

Key words. economic lot-sizing, polyhedral description, plant location formulation, separation

AMS subject classification. 90B

1. Introduction. Good integer linear programming formulations for NP-hard problems are a valuable aid in solving these problems with linear-programming-based solution methods. One way of improving formulations with strong valid inequalities is by looking at relaxations or substructures that are polynomially solvable. For such a relaxation, contrary to the original NP-hard problem, we may be able to find a complete linear description. This holds, for instance, for several economic lot-sizing problems. Research on strong valid inequalities for the uncapacitated single-item economic lot-sizing problem (ELS), as defined by Wagner and Whitin [17] and Manne [11], started with Barany, Van Roy, and Wolsey [1], [2], who developed the so-called (l, S)-inequalities. The polyhedral structure of generalizations of the economic lot-sizing problem has been the subject of several papers. Leung, Magnanti, and Vacani [9] and Pochet and Wolsey [13] give strong valid inequalities for the capacitated ELS. Pochet and Wolsey [12] consider the extension with backlogging. They describe an implicit characterization of the problem.

We consider the extension of the economic lot-sizing problem with start-up costs included (ELSS), i.e., costs for switching on a machine or changing over between different items. Schrage [15] introduced these costs to distinguish between normal set-up costs, which are incurred in each period that a certain item is produced, and costs that appear only in the first of a consecutive set of periods in which an item is produced. Problems in which start-up costs appear have been studied by Van Wassenhove and Vanderhenst [16], Karmarkar and Schrage [7] and Fleischmann [4]. The standard dynamic programming formulation of the economic lot-sizing problem with start-up costs can be solved in $O(T\log T)$ time; here T is the length of the planning horizon (see van Hoesel [5]).

The polyhedral structure of several mixed integer programming formulations for ELSS was first investigated by Wolsey [18]. For the formulation in a natural set of variables, he derived a class of strong valid inequalities by generalizing the (l, S)-inequalities for the corresponding formulation of the economic lot-sizing problem (Barany, Van Roy, and Wolsey [2]). In this manuscript, we further generalize these inequalities to the so-called (l, R, S)-inequalities. The main result that we present is that these inequalities imply a complete linear description for this formulation. We use a proof technique due to Lovász [10], which appears to be especially suitable for problems where a greedy al-

*Received by the editors March 6, 1991; accepted for publication (in revised form) September 24, 1992.

†Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands.

‡Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands.

§Center of Operations Research and Econometrics, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium.

gorithm solves the dual linear program arising from a complete linear description of the problem (see van Hoesel, Wagelmans, and Kolen [6]). Conditions under which the (l, R, S)-inequalities are facet-defining are provided by van Hoesel [5]). In addition, we discuss separation for the (l, R, S)-inequalities by formulating this problem as a set of T shortest path problems on acyclic networks, each with $O(T^2)$ nodes.

A related formulation for ELS is the plant location reformulation, in which the production variables are split. This formulation has been introduced by Krarup and Bilde [8]. The plant location reformulation for ELSS is shown to be at least as strong as the formulation in the original variables. This is done by viewing the inequalities as di-cut inequalities (Rardin and Wolsey [14]) in a fixed-charge min-cost flow problem. For other related formulations and similar results, see Wolsey [18] and Eppen and Martin [3].

In §2 ELSS is formulated as a mixed integer programming problem. The (l, R, S)-inequalities are introduced and shown to be valid. In §3 it is shown that the (l, R, S)-inequalities provide a complete linear description. The separation algorithm for the (l, R, S)-inequalities can also be found in this section. In §4 the plant location model is discussed. Finally, in §5 some concluding remarks are made.

2. Formulation of ELSS: The (*l, R, S*)-inequalities. Consider ELSS with a planning horizon consisting of T periods. For each period t, a demand d_t must be satisfied by production in one or more of the periods in $\{1, \ldots, t\}$. The costs for production in period t are c_t per unit. If production takes place in period t, a set-up must be performed at a cost of f_t. This is the formulation of ELS as defined by Wagner and Whitin [17]. In ELSS there are additional fixed costs for start-ups: Each set of consecutive periods in which a set-up is performed should begin with a period t in which a start-up is performed at a cost of g_t. For reasons of simplicity, in the formulation we present, the inventory variables are deleted (see, for instance, Wolsey [18] or van Hoesel [5]). ELSS can be modelled as a mixed integer program with the following parameters and variables.

Parameters:
d_t $(1 \le t \le T)$: the demand of the item in period t;
c_t $(1 \le t \le T)$: the unit production cost of the item in period t;
f_t $(1 \le t \le T)$: the set-up cost of the item in period t;
g_t $(1 \le t \le T)$: the start-up cost of the item in period t;
Variables:
x_t $(1 \le t \le T)$: the production of the item in period t;

$$y_t\ (1 \le t \le T) \begin{cases} 1 & \text{if a set-up of the item is incurred in period } t; \\ 0 & \text{otherwise.} \end{cases}$$

$$z_t\ (1 \le t \le T) \begin{cases} 1 & \text{if a start-up of the item is incurred in period } t; \\ 0 & \text{otherwise;} \end{cases}$$

$$(1) \qquad \text{(ELSS)} \quad \min \sum_{t=1}^{T} (g_t z_t + f_t y_t + c_t x_t),$$

$$(2) \qquad \text{s.t.} \quad \sum_{t=1}^{T} x_t = d_{1,T},$$

$$(3) \qquad \sum_{\tau=1}^{t} x_\tau \ge d_{1,t}, \qquad (1 \le t \le T-1),$$

$$y_t \le y_{t-1} + z_t, \qquad (y_0 \equiv 0), \qquad (1 \le t \le T), \tag{4}$$

$$x_t \le d_{t,T}y_t, \qquad (1 \le t \le T), \tag{5}$$

$$x_t \ge 0, \qquad (1 \le t \le T), \tag{6}$$

$$y_t, z_t \in \{0,1\}, \qquad (1 \le t \le T). \tag{7}$$

By $d_{s,t}$ $(1 \le s \le t \le T)$, we denote the cumulative demand of the periods $\{s, \dots, t\}$, i.e., $d_{s,t} = \sum_{\tau=s}^{t} d_\tau$.

Constraint (2) restricts production to the total demand over the planning horizon. Constraints (3) ensure that ending inventory in each period is nonnegative. Constraints (4) model the start-ups. Constraints (5) force a set-up in a period with positive production. Note that (2) and (3) (for $t-1$) imply the upper bound $d_{t,T}$ on the production in each period t as mentioned in (5).

The remainder of this section is devoted to the description of the (l, R, S)-inequalities and a proof of their validity. Take an arbitrary period $l \in \{1, \dots, T\}$ and let $N_l = \{1, \dots, l\}$. Let S be an arbitrary subset of $\{1, \dots, N_l\}$ and R be a subset of S, such that the first element in S is also in R. The corresponding (l, R, S)-inequality is defined as follows:

$$\sum_{t \in N_l \setminus S} x_t + \sum_{t \in R} d_{t,l}y_t + \sum_{t \in S \setminus R} d_{t,l}(z_{p(t)+1} + \dots + z_t) \ge d_{1,l}, \tag{8}$$

where $p(t) \equiv \max\{j \in S | j < t\}$. If $S \cap \{1, \dots, t-1\} = \emptyset$, then $p(t) \equiv 0$.

Example. $l = 15; S = \{2,4,5,7,10,11,14\}; R = \{2,10\}$. The coefficients of the left-hand side of the (l, R, S)-inequality are given in the following table:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_t	1		1			1		1	1			1	1		1
y_t		$d_{2,l}$								$d_{10,l}$					
z_t			$d_{4,l}$	$d_{4,l}$	$d_{5,l}$	$d_{7,l}$	$d_{7,l}$				$d_{11,l}$	$d_{14,l}$	$d_{14,l}$	$d_{14,l}$	

The inequalities derived in Wolsey [18] are a special case of the (l, R, S)-inequalities in the sense that each maximal set of consecutive periods in S should begin with a period in R there. The above example is not included, because the set $\{4, 5\}$ does not begin with an element in R $(4 \in S \setminus R)$.

LEMMA 2.1. *The* (l, R, S)*-inequalities are valid.*

Proof. Take an arbitrary (l, R, S)-inequality as defined above. Denote an arbitrary feasible solution by $(x, y, z) = \{x_t, y_t, z_t | t = 1, \dots, T\}$. We prove that this solution satisfies the (l, R, S)-inequality. We distinguish two cases.

Case 1. S does not contain a period with positive production, i.e., $x_t = 0$ for all $t \in S$. Then

$$\sum_{t \in N_l \setminus S} x_t = \sum_{t=1}^{l} x_t \ge d_{1,l}.$$

Case 2. S contains a period with production. Let s be the first such period, i.e., $x_s > 0$. First,

$$\sum_{t \in N_l \setminus S} x_t \geq \sum_{t \in N_{s-1} \setminus S} x_t = \sum_{t \in N_{s-1}} x_t \geq d_{1,s-1}.$$

Now choose τ as small as possible such there are set-ups in all periods $\{\tau, \dots, s\}$, i.e., $z_\tau = y_\tau = \cdots = y_s = 1$. Since at least one of these variables appears in the left-hand side of the (l, R, S)-inequality with a coefficient that is at least $d_{s,l}$, the inequality is satisfied by this solution. □

3. Linear description of ELSS and separation for the (*l*, *R*, *S*)-inequalities. The main result in this section is that addition of the (l, R, S)-inequalities to the model for ELSS gives a complete polyhedral description. More precisely, we show the following result.

THEOREM 3.1. *Constraints* (2), (4), (6), *the variable bounding constraints* $0 \leq y_t, z_t \leq 1$ $(1 \leq t \leq T)$, *and the* (l, R, S)*-inequalities* (8) *describe the convex hull of* ELSS.

Note that inequalities (3) are $(t, \emptyset, \emptyset)$-inequalities. Inequalities (5) can be derived from (2) and (8), where $S = R = \{t\}$ and $l = T$.

The technique we use to prove the theorem is somewhat different from the usual techniques. Basically, the idea is to show that, for an arbitrary objective function, denoted by $\sum_{t=1}^{T}(\alpha_t x_t + \beta_t y_t + \gamma_t z_t)$, the set of optimal solutions, denoted by $M(\alpha, \beta, \gamma)$, satisfies one of the inequality constraints as equality. Clearly, then the inequality constraints must include all facets of the convex hull of solutions.

We consider an arbitrary cost function $\sum_{t=1}^{T}(\alpha_t x_t + \beta_t y_t + \gamma_t z_t)$ and we denote its set of optimal solutions by $M(\alpha, \beta, \gamma)$.

Case 0. $\min\{\alpha_t | t = 1, \dots, T\} = \delta \neq 0$.

As $\sum_{t=1}^{T} x_t = d_{1,T}$, we can subtract δ times inequality (2) from the objective function without changing the set of optimal solutions.

Thus, in the following, we can assume that $\min\{\alpha_t | t = 1, \dots, T\} = 0$.

Case 1. $\gamma_t < 0$ for some $t \in \{1, \dots, T\}$.

Any solution with $z_t = 0$ can be improved by setting $z_t = 1$. Thus $M(\alpha, \beta, \gamma) \subseteq \{(x, y, z) | z_t = 1\}$.

Case 2. $\gamma_t \geq 0$ for all t, $\beta_t < 0$ for some t.

(i) $\beta_t + \gamma_t < 0$ for some t.

Any solution with $y_t = 0$ can be improved by setting $y_t = z_t = 1$. Thus $M(\alpha, \beta, \gamma) \subseteq \{(x, y, z) | y_t = 1\}$.

(ii) $\beta_t + \gamma_t \geq 0$ for all t.

Let $s = \min\{t | \beta_t < 0\}$. We show that $M(\alpha, \beta, \gamma) \subseteq \{(x, y, z) | y_{s-1} + z_s = y_s\}$. As $\gamma_s > 0$, any solution with $y_{s-1} = z_s = 1$ can be improved by setting $z_s = 0$. In addition, any solution with $y_{s-1} + z_s = 1$ and $y_s = 0$ can be improved by setting $y_s = 1$. Thus, the claim follows.

We are left with objective functions satisfying $\min\{\alpha_t | t = 1, \dots, T\} = 0$; $\beta_t \geq 0$ $(t = 1, \dots, T)$; $\gamma_t \geq 0$ $(t = 1, \dots, T)$. In the following, l, S, and R are determined, step by step, in this order. First, the period l is fixed. This is done such that the demands of the periods $\{l+1, \dots, T\}$ can be produced at no cost, with respect to the given objective function. For notational convenience, a period $T+1$ is defined with $\alpha_{T+1} = 0$, $\beta_{T+1} = 0$, $\gamma_{T+1} = 0$, and the demand of $T + 1$ is supposed to be positive.

Choice of l. Take k and m ($1 \leq k \leq m \leq T+1$), both minimal, such that $\gamma_k = \beta_k = \cdots = \beta_m = \alpha_m = 0$. The period l is chosen as the last period in $\{1, \ldots, m-1\}$ with positive demand, i.e., $d_l > 0$, and $d_{l+1} = \cdots = d_{m-1} = 0$. If $d_{1,m-1} = 0$, then $l \equiv 0$.

Case 3. Since $d_{l+1,m-1} = 0$, any production in the periods $\{l+1, \ldots, T\}$ can be moved to period m at no cost, by taking $z_k = y_k = \ldots = y_m = 1$.

It follows that (i) if $\alpha_t > 0$ for some $t \geq l+1$, then $M(\alpha,\beta,\gamma) \subseteq \{(x,y,z)|x_t = 0\}$; (ii) If $\beta_t > 0$ for some $t \geq \min\{k, l+1\}$, then $M(\alpha,\beta,\gamma) \subseteq \{(x,y,z)|y_t = 0\}$; (iii) If $\gamma_t > 0$ for some $t \geq \min\{k, l+1\}$, then $M(\alpha,\beta,\gamma) \subseteq \{(x,y,z)|z_t = 0\}$.

If the contrary holds in one of these cases, then the solution can be improved by taking the solution mentioned above.

Note that, if $l = 0$, then we are finished. In that case, the only objective function left is the one with zero cost coefficients for all variables. Thus, in the following, we may assume that $l > 0$. Moreover, we may suppose that, for $t \geq \min\{k, l+1\}$, we have $\beta_t = \gamma_t = 0$, and, for $t \geq l+1$, we have $\alpha_t = 0$. We proceed by specifying the choices of S and R.

Choice of S. $S := \{t \leq l | \alpha_t = 0\}$.

Choice of R. $R := \{t \in S | \beta_t > 0\}$.

With regard to the following case note that, if $l = T$, then S is not empty, since $\min\{\alpha_t | t = 1, \ldots, T\} = 0$.

Case 4. If $S = \emptyset$, then $M(\alpha,\beta,\gamma) \subseteq \{(x,y,z)| \sum_{t=1}^{l} x_t = d_{1,l}\}$.

$S = \emptyset$ implies that $l < T$ and $\alpha_t > 0$ for $t \in \{1, \ldots, l\}$. Therefore, if $I_l > 0$ (the inventory at the end of period l), then the production costs can be reduced by transferring units of production from the last production period in $\{1, \ldots, l\}$ to period $l+1$. Note that $\alpha_{l+1} = \beta_{l+1} = \gamma_{l+1} = 0$. This proves the claim, since $I_l = 0$ implies that $\sum_{t=1}^{l} x_t = d_{1,l}$.

In the following, we may suppose that $S \neq \emptyset$.

Case 5. Suppose that, for some $t \in S \backslash R$, there is an $s \in \{p(t)+1, \ldots, t\}$ with $\beta_s > 0$. Then for this t the following holds: $M(\alpha,\beta,\gamma) \subseteq \{(x,y,z)|y_s + z_{s+1} + \cdots + z_t = y_t\}$.

Note that the equation $y_s + z_{s+1} + \cdots + z_t = y_t$ implies that $y_{t-1} + z_t = y_t$ ($s < t$ since $t \notin R$). Before we proceed, some properties of the coefficients of the variables are stated. First, $\beta_s > 0$ and $\beta_{s+1}, \ldots, \beta_t = 0$ (since $t \notin R$). Second, $\alpha_s, \ldots, \alpha_{t-1} > 0$, since $s \in \{p(t)+1, \ldots, t-1\}$, and, since $t \in S$, we have $\alpha_t = 0$. Finally, if $\gamma_\tau = 0$ for some $\tau \in \{s+1, \ldots, t\}$, then $\gamma_\tau = \beta_\tau = \cdots = \beta_t = \alpha_t = 0$, which implies that $l < t$, a contradiction. It follows that $\gamma_{s+1}, \ldots, \gamma_t > 0$. Concluding, the variables on the left-hand side of the constraint $y_s + z_{s+1} + \cdots + z_t = y_t$ have positive cost coefficients, and the variable on the right-hand side has cost coefficient zero.

(i) First, suppose that $y_s + z_{s+1} + \cdots + z_t \geq 2$. Let u be the first period in $\{s+1, \ldots, t\}$ in which the variable z_u has value 1. If $y_s = 1$, then a cost reduction can be achieved by setting $y_{s+1}, \ldots, y_u = 1$, and $z_u = 0$. If $y_s = 0$, then let v be the first period after u with $z_v = 1$. In this case, a cost reduction can be obtained by setting $y_u = \cdots = y_v = 1$ and $z_v = 0$.

(ii) Second, suppose that $y_s + z_{s+1} + \cdots + z_t = 1$ and $y_t = 0$. If no production takes place in $s, \ldots, t-1$, then setting $y_s + z_{s+1} + \cdots + z_t = 0$ leads to a cost reduction without losing feasibility. Otherwise, production takes place in one or more of the periods $\{s, \ldots, t-1\}$. Let u be the last production period in $\{s, \ldots, t-1\}$. If $I_{t-1} > 0$, then production from period u can be transferred to t, at a cost reduction, since $\beta_{u+1} = \cdots = \beta_t = \alpha_t = 0$, and $\alpha_u, \ldots, \alpha_{t-1} > 0$. Thus, we may assume that $I_{t-1} = 0$. Then d_t must be zero, since $y_t = 0$, which implies that $l > t$. Thus, the positive demand $d_{t+1,l}$ is produced at a positive cost in some of the periods in the set $\{t+1, \ldots, l\}$; otherwise, l would have been chosen smaller (see choice of l). Now, however, it is cheaper to produce this demand in period t at zero cost. Of course, this implies that $y_{u+1} = \cdots = y_t = 1$, where u is the last production period in $\{s, \ldots, t-1\}$.

In the following, we may assume that, for each $t \in S \backslash R$, we have $\beta_{p(t)+1} = \cdots = \beta_t = 0$. The following case treats the problem, where the first element in S in not in R.

Case 6. If $\beta_1 = \cdots = \beta_t = 0$, where t is the first element in S, then we have the following: $M(\alpha, \beta, \gamma) \subseteq \{(x, y, z) | \sum_{\tau=1}^{t-1} x_t = d_{1,t-1}\}$.

By definition of t, we have $\{1, \ldots, t-1\} \cap S = \emptyset$, and thus $\alpha_1, \ldots, \alpha_{t-1} > 0$. Moreover, since $t \in S$, we have $\alpha_t = 0$. If $y_1 + \cdots + y_t = 0$, then $\sum_{\tau=1}^{t} x_t = 0 = d_{1,t-1}$. If $y_1 + \cdots + y_t \geq 1$, then we can produce in t at no additional production costs, and thus, since $\alpha_1, \ldots, \alpha_{t-1} > 0$, we have $I_{t-1} = 0$. Therefore, $\sum_{\tau=1}^{t-1} x_t = d_{1,t-1}$.

It follows from Case 6 that we may assume that the first period in S is in R.

Case 7. We claim that the (l, R, S)-inequality

$$(9) \qquad \sum_{t \in N_l \backslash S} x_t + \sum_{t \in R} d_{t,l} y_t + \sum_{t \in S \backslash R} d_{t,l}(z_{p(t)+1} + \cdots + z_t) \geq d_{1,l}$$

is satisfied as equality for all points in $M(\alpha, \beta, \gamma)$.

The following properties concerning the cost coefficients are valid:

For $t \in N_l \backslash S$, $\alpha_t > 0$, by definition of S;

For $t \in R$, $\beta_t > 0$, $\alpha_t = 0$, by definition of R;

For $t \in S \backslash R$, $\gamma_{p(t)+1}, \ldots, \gamma_t > 0$, since $\beta_{p(t)+1} = \cdots = \beta_t = \alpha_t = 0$. Otherwise, $l < t$.

(i) Suppose that $y_t = 0$ for all $t \in R$ and $z_{p(t)+1} + \cdots + z_t = 0$ for all $t \in S \backslash R$. Then $y_t = 0$ for all $t \in S$, which implies that there is no production in the periods of S. Moreover, the contribution of the variables in the left-hand side of the periods in S is zero. It remains to prove that $\sum_{t \in N_l \backslash S} x_t = d_{1,l}$. If $I_l > 0$, then the costs can be decreased by transferring production from the last production period in $\{1, \ldots, l\}$ to $l+1$. Recall that $\alpha_{l+1} = \beta_{l+1} = \gamma_{l+1} = 0$. Thus $I_l = 0$. This proves the claim.

Now take the minimal $t \in S$ such that $y_t = 1$ (if $t \in R$) or $z_{p(t)+1} + \cdots + z_t \geq 1$ (if $t \in S \backslash R$).

(ii) If $t \in R$, then by the minimality of t this gives the following: (a) For $\tau < t$, $\tau \in R$, $y_\tau = 0$, and (b) For $\tau < t$, $\tau \in S \backslash R$, $z_{p(\tau)+1}, \ldots, z_\tau = 0$. Since $\alpha_t = 0$, and $y_t = 1$, the production of $d_{t,l}$ can take place at no additional cost in t. Therefore, the following hold:

(c) For $\tau > t$, $\tau \in N_l \backslash S$, $x_\tau = 0$, since $\alpha_\tau > 0$,

(d) For $\tau > t$, $\tau \in R$, $y_\tau = 0$, since $\beta_\tau > 0$,

(e) For $\tau > t$, $\tau \in S \backslash R$, $z_{p(\tau)+1} = \cdots = z_\tau = 0$, since $\gamma_{p(\tau)+1}, \ldots, \gamma_\tau > 0$.

Thus, the value of the left-hand side of the (l, R, S)-inequality is reduced to the quantity $\sum_{\tau \in N_{t-1} \backslash S} x_\tau + d_{t,l}$. It remains to be proved that $\sum_{\tau \in N_{t-1} \backslash S} x_\tau$ equals $d_{1,t-1}$. Suppose that $I_{t-1} > 0$. From the minimality of t, it follows that there are no production periods in $S \cap N_{t-1}$, and therefore we can reduce the production costs by transferring production from the last production period in N_{t-1} to t.

(iii) If $t \in S \backslash R$, then let $s \in \{p(t)+1, \ldots, t\}$ be the first period with $z_s = 1$. By the minimality of t we have the following:

(a) For $\tau < t$, $\tau \in R$, $y_\tau = 0$,

(b) For $\tau < t$, $\tau \in S \backslash R$, $z_{p(\tau)+1}, \ldots, z_\tau = 0$.

Moreover, since t is minimal $\{s, \ldots, t-1\} \cap S = \emptyset$, and therefore $p(t) < s$. Since $\beta_s = \cdots = \beta_t = \alpha_t = 0$, the production of $d_{t,l}$ can take place at no additional cost in t. Therefore

(c) For $\tau > t$, $\tau \in N_l \backslash S$, $x_\tau = 0$, since $\alpha_\tau > 0$,
(d) For $\tau > t$, $\tau \in R$, $y_\tau = 0$, since $\beta_\tau > 0$,
(e) For $\tau > t$, $\tau \in S \backslash R$, $z_{p(\tau)+1} = \cdots = z_\tau = 0$, since $\gamma_{p(\tau)+1}, \ldots, \gamma_\tau > 0$.

Note that $z_{s+1}, \ldots, z_t = 0$; otherwise, the costs can be reduced by setting $y_{s+1} = \cdots = y_t = 1$. By the minimality of s, it follows that $z_{p(t)+1} = \cdots = z_{s-1} = 0$.

Since the coefficient of z_s is $d_{t,l}$, the value of the left-hand side of the (l, R, S)-inequality is reduced to the quantity $\sum_{\tau \in N_{t-1} \backslash S} x_t + d_{t,l}$. It remains to be proved that $\sum_{\tau \in N_{t-1} \backslash S} x_t$ equals $d_{1,t-1}$. Suppose that $I_{t-1} > 0$. From the minimality of t, it follows that there are no production periods in $S \cap N_{t-1}$, and therefore we can reduce the production costs by transferring production from the last production period in N_{t-1} to t.

This ends the proof of the theorem. The following theorem shows which of the (l, R, S)-inequalities define facets of the convex hull of ELSS (for a proof, see van Hoesel [5]).

THEOREM 3.2. *Suppose that the demands d_t $(1 \le t \le T)$ are positive. The (l, R, S)-inequalities define facets, if and only if the following conditions hold.* $1 \in S \neq \emptyset$ *and* $l < T$ *or* $|R| = 1$.

A separation algorithm for the *(l, **R**, S)*-inequalities. Here, we show that the separation algorithm for the (l, R, S)-inequalities can be formulated as a shortest path problem.

We fix l. Then we define the following three nodes for each $t \in \{0, ..., l\}$: u_t, v_t, and w_t. Moreover, a starting node n_0 and an ending node n_l are defined. There are the following three arcs with n_0 as a tail: (n_0, u_0), (n_0, v_0), and (n_0, w_0) all with zero costs. Moreover, there are the following three arcs with head n_l: (u_l, n_l), (v_l, n_l), and (w_l, n_l), also with zero costs. To model the (l, R, S)-inequalities in a network, we define three types of arcs, below:

Type 1. arcs (u_{t-1}, u_t), (v_{t-1}, u_t), (w_{t-1}, u_t) with cost x_t,
Type 2. arcs (u_{t-1}, v_t), (v_{t-1}, v_t), (w_{t-1}, v_t) with cost $d_{t,l} y_t$,
Type 3. arcs $(v_{p(t)}, w_t)$, $(w_{p(t)}, w_t)$ with cost $\sum_{\tau=p(t)+1}^{t-1} x_\tau + \sum_{\tau=p(t)+1}^{t} d_{t,l} z_\tau$.

Each path in the network corresponds to the left-hand side of a unique (l, R, S)-inequality. In particular, the nodes $\{v_t\}$ and $\{w_t\}$ correspond to the sets R and $S \backslash R$, respectively. Therefore, the shortest path in the network can be compared with $d_{1,l}$ to find a violated (l, R, S)-inequality for a fixed l. See Fig. 1.

There are $O(l^2)$ arcs in the network, and, since it is acyclic, the shortest path problem in the network can be solved in $O(l^2)$ time. Repeated for each period $l \in \{1, \ldots, T\}$, this leads to an $O(T^3)$ algorithm to find the most-violated (l, R, S)-inequality. This is to be compared with the single max-flow calculation on a graph with $O(T^3)$ nodes derived in Rardin and Wolsey [14].

4. The uncapacitated facility location reformulation of ELSS. In this section, we consider the uncapacitated facility location reformulation of ELSS (ELSS-UFL) in which the production variables are split. The variables $x_{t,\tau}$ $(1 \le t \le \tau \le T)$ are introduced as the production of the item in period t to satisfy demand in period τ. Clearly, the connection with the original production variables is $x_t = \sum_{\tau=t}^{T} x_{t,\tau}$. ELSS-UFL is modeled as follows:

$$(10) \qquad \text{(ELSS-UFL)} \qquad \min \sum_{t=1}^{T} \left(g_t z_t + f_t y_t + c_t \left(\sum_{\tau=t}^{T} x_{t,\tau} \right) \right),$$

$$(11) \qquad \text{s.t.} \quad \sum_{t=1}^{\tau} x_{t,\tau} = d_\tau, \qquad (1 \le \tau \le T),]$$

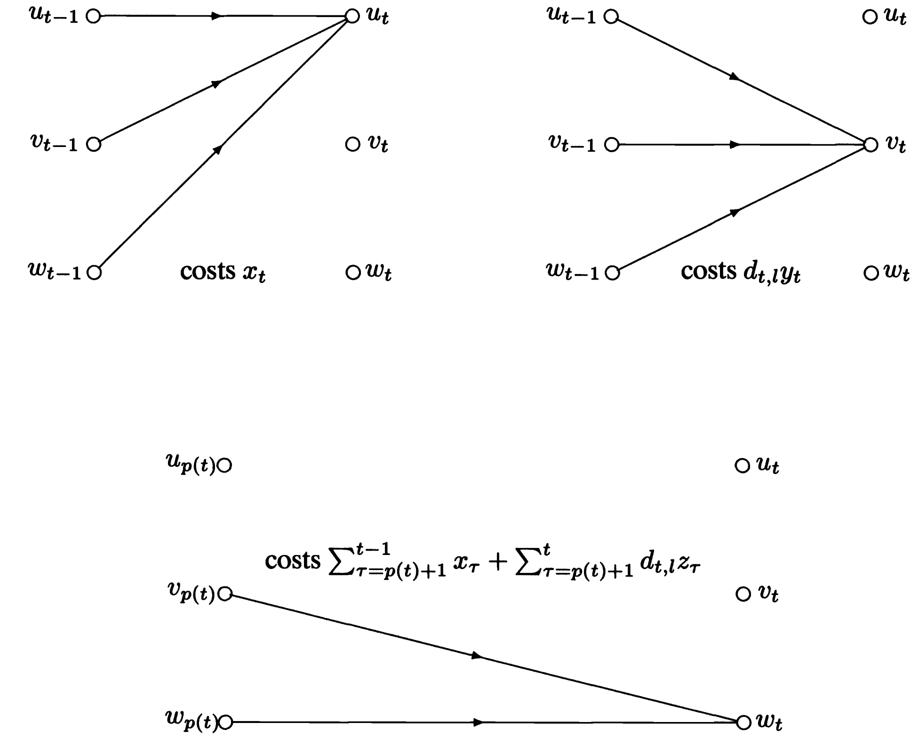


FIG. 1. *Arcs in the network for the separation problem.*

$$y_t \le y_{t-1} + z_t, \qquad (y_0 \equiv 0) \qquad (1 \le t \le T), \tag{12}$$

$$x_{t,\tau} \le d_\tau y_t, \qquad (1 \le t \le \tau \le T), \tag{13}$$

$$x_{t,\tau} \ge 0, \qquad (1 \le t \le \tau \le T), \tag{14}$$

$$y_t, z_t \in \{0,1\}, \qquad (1 \le t \le T). \tag{15}$$

The LP-relaxation of ELSS-UFL is not tight, in the sense that it still allows fractional solutions. By adding the following constraints, the so-called (r, s, τ)-inequalities, we get a reformulation of ELSS, which is at least as strong as the formulation given in the previous section.

Let $1 \le r \le s \le \tau \le T$,

$$\sum_{t \in N_\tau \setminus \{r,\dots,s\}} x_{t,\tau} + d_\tau(y_r + z_{r+1} + \ldots + z_s) \ge d_\tau$$

These inequalities can be found in Wolsey [18]. The (r, s, τ)-inequalities can be viewed as cuts in the following fixed-charge multicommodity network flow problem. The flow network consists of the following vertices and arcs. There is a source s, and there are the following three layers of nodes: $\{u_t | 1 \le t \le T\}$, $\{v_t | 1 \le t \le T\}$, and $\{w_t | 1 \le t \le T\}$.

The arcs (s, u_t), $(1 \le t \le T)$ model the fixed start-up charges; i.e., if such an arc contains a positive flow, then $z_t = 1$.

The arcs (u_t, v_t), $(1 \le t \le T)$ model the fixed set-up charges.

The arcs (v_t, u_{t+1}), $(1 \le t \le T-1)$ are included to allow for multiple set-ups in consecutive periods without a start-up in these periods.

The arcs (v_t, w_t), $(1 \le t \le T)$, model the production in period t.

The production is exactly equal to the flow through the arc. Finally, the flows through the arcs (w_t, w_{t+1}), $(1 \le t \le T-1)$ denote the inventory at the end of period t or, equivalently, at the beginning of period $t+1$. There are T different commodities τ, $(1 \le \tau \le T)$ in the network, consisting of the source s and sink w_τ, and with a demand of d_τ units of flow. For a commodity τ, the flow through the arcs (v_t, w_t), $(1 \le t \le \tau)$ is the production in period t for period τ, and therefore it is denoted by $x_{t,\tau}$. The arcs (w_t, w_{t+1}), $(1 \le t \le \tau - 1)$ contain the inventory at the end of period t for demand in period τ. It is denoted by $I_{t,\tau}$.

Example. Consider a feasible flow of d_τ units of a given commodity τ. The flow "through" a cut (V, W) separating s and w_τ is at least d_τ units. For each type of arc, we can derive an upper bound on the flow through such arcs as follows. The flow through an arc (s, u_t) is bounded by $d_\tau z_t$, the flow through an arc (u_t, v_t) is bounded by $d_\tau y_t$. Finally, the flow through an arc (v_t, w_t) equals $x_{t,\tau}$. We consider cuts that "contain" these types of arcs only. In that case, it follows trivially that the sum of these upper bounds for all arcs in the cut, constitutes an upper bound on the flow d_τ. See Fig. 2.

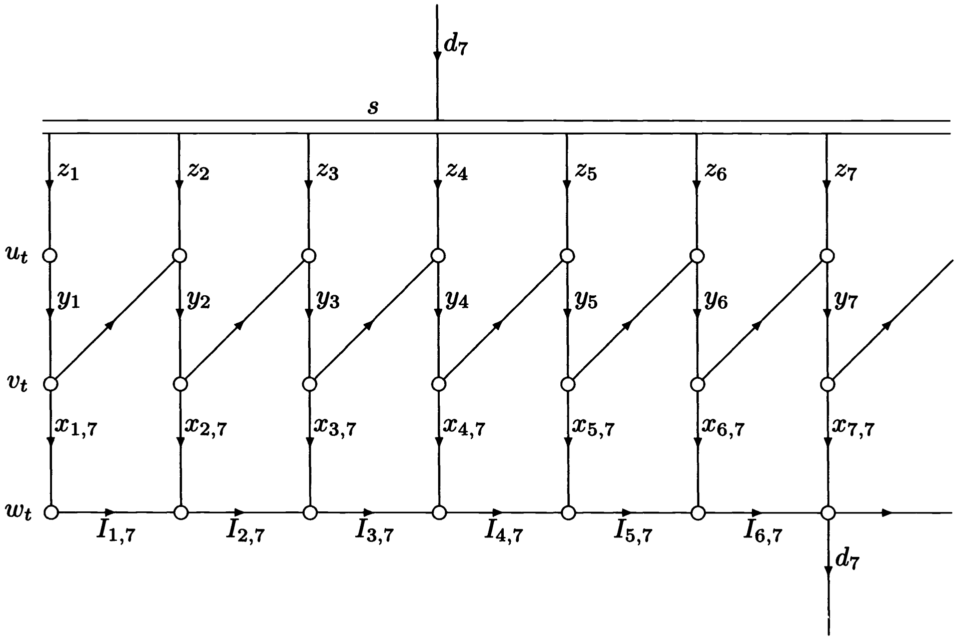


FIG. 2. *Fixed-charge flow network.*

The cut (V, W) that constitutes the validity of the (r, s, τ)-inequality is the following:

$$V = \{s, u_1, \dots, u_r, u_{s+1}, \dots, u_\tau, v_1, \dots, v_{r-1}, v_{s+1}, \dots, v_\tau\};$$

$$W = \{u_{r+1}, \dots, u_s, v_r, \dots, v_s, w_1, \dots, w_\tau\}.$$

Clearly, we may take more closed intervals $\{r, \dots, s\}$, as long as they are mutually disjoint. In fact, for a given l, R, and S, we take the maximal closed intervals that start with a period in R and end with a period in $S\backslash R$ such that no intermediate periods are in R. This leads to the following valid inequality for τ:

$$\sum_{t\in N_\tau\backslash S_\tau} x_{t,\tau} + \sum_{t\in R_\tau} d_\tau y_t + \sum_{t\in S_\tau\backslash R_\tau} d_\tau(z_{p(t)+1} + \cdots + z_t) \geq d_\tau$$

Here $R_\tau = N_\tau \cap R$ and $S_\tau = N_\tau \cap S$. Addition of these inequalities for $\tau = 1, \ldots, l$ results in the (l, R, S)-inequality. This result can be found in Rardin and Wolsey [14]. There it has been shown that the (r, s, τ)- inequalities are so-called di-cut inequalities in the uncapacitated fixed-charge network flow model.

Note. If the demands in all periods are positive, we can obtain a more compact model than ELSS-UFL. The latter contains $O(T^3)$ constraints. This can be reduced to $O(T^2)$ as follows. ELSS has always an optimal solution in which $x_{t,\tau}/d_\tau$ are nonincreasing in τ (t fixed). A simple exchange argument then shows that the (r, s, s)-inequalities suffice.

5. Concluding remarks. We characterized the convex hull of the set of feasible solutions of ELSS by use of the (l, R, S)-inequalities and we provided a separation algorithm for these inequalities. This formulation has $O(T)$ variables and an exponentional number of constraints. In addition, we showed that the uncapacitated facility location reformulation is tight. This formulation has $O(T^2)$ variables and $O(T^3)$ constraints. If the demands are strictly positive, the number of constraints can be reduced to $O(T^2)$.

These results are of a purely theoretical nature. It is therefore hard to judge which of the two formulations is better used in practical problems in which ELSS appears as a relaxation.

REFERENCES

[1] I. Barany, T. J. Van Roy, and L. A. Wolsey, *Strong formulations for multi-item capacitated lotsizing*, Management Sci., 30 (1984), pp. 1255–1261.

[2] ———, *Uncapacitated lot-sizing: the convex hull of solutions*, Math. Programming Study, 22 (1984), pp. 32–43.

[3] G. D. Eppen and R. K. Martin, *Solving multi-item capacitated lot-sizing problems using variable redefinition*, Oper. Res., 35 (1987), pp. 268–277.

[4] B. Fleischmann, *The discrete lot-sizing and scheduling problem*, European J. Oper. Res., 44 (1990), pp. 337–348.

[5] S. van Hoesel, *Models and Algorithms for Single-Item Lot Sizing Problems*, Ph.D. thesis, Econometric Institute, Erasmus University Rotterdam, the Netherlands, 1991.

[6] S. van Hoesel, A. Wagelmans, and A. Kolen, *A dual algorithm for the economic lot-sizing problem*, European J. Oper. Res., 52 (1991) pp. 315–325.

[7] U. S. Karmarkar and L. Schrage, *The deterministic dynamic product cycling problem*, Oper. Res., 33 (1985), pp. 326–345.

[8] J. Krarup and O. Bilde, *Plant location, set covering, and economic lot-size: An $o(mn)$ algorithm for structured problems*, Numerische Methoden bei Optimierungsaufgaben, Band 3: Optimierung bei Graphentheorietischen und Ganzzahligen Problemen, Birkhäuser, Boston, Basel, 1977, pp. 155–186.

[9] J. M. Y. Leung, T. L. Magnanti, and R. Vachani, *Facets and algorithms for capacitated lotsizing*, Math. Programming, 45 (1989), pp. 331–359.

[10] L. Lovász, *Graph theory and integer programming*, Ann. Discrete Math., 4 (1979), pp. 141–158.

[11] A. S. Manne, *Programming of economic lot sizes*, Management Sci., 4 (1958), pp. 115–135.

[12] Y. Pochet and L. A. Wolsey, *Lot-size models with backlogging: strong reformulations and cutting planes*, Math. Programming, 40 (1988), pp. 317–335.

[13] ———, *Solving multi-item lot-sizing problems using strong cutting planes*, Management Sci., 37 (1991), pp. 53–67.

[14] R. L. Rardin and L. A. Wolsey, *Valid inequalities and projecting the multicommodity extended formulation for uncapacitated fixed charge network flow problems*, CORE Discussion Paper 9024, Center for Operations Research and Econometrics, Louvain-la-Neuve, Belgium, 1990.

[15] L. Schrage, *The multiproduct lot scheduling problem*, Deterministic and Stochastic Scheduling, M.A.H. Dempster, J.K. Lenstra and A.H.G. Rinnooy Kan, eds., Nato advanced Study Institutes Series, D. Reidel, Holland, 1984.

[16] L.N. Van Wassenhove and P. Vanderhenst, *Planning production in a bottleneck department*, European J. Oper. Res., 12 (1983), pp. 127–137.

[17] H. W. Wagner and T. H. Whitin, *Dynamic version of the economic lot size model*, Management Sci., 1 (1958), pp. 88–96.

[18] L. A. Wolsey, *Uncapacitated lot-sizing problems with start-up costs*, Oper. Res., 37 (1989), pp. 741–747.

SIAM J. DISCRETE MATH
Vol. 7, No. 1, pp. 152–155, February 1994

015

A NOTE ON MULTISET PERMUTATIONS*

LILY YEN†

Abstract. The author studies permutations of the multiset $\{1,1,2,2,\ldots,m,m,m+1,m+2,\ldots,n\}$ such that $1,2,\ldots,n$ occurs as a not-necessarily consecutive subsequence. From the theory of symmetric functions, the generating function for the number of these permutations is known [Goulden and Jackson, *Combinatorial Enumeration*, John Wiley, New York, 1983, p. 73]. It is used to obtain a recurrence relation and then to give a purely combinatorial proof of the recurrence.

Key words. permutations, multiset, bijection, recurrence

AMS subject classifications. 05A19, 05E05

1. Introduction. We want to give a combinatorial proof of the following theorem.

THEOREM. *Let $\mathcal{I}_m(n)$ be the set of permutations of the multiset*

$$\{1,1,2,2,\ldots,m,m,m+1,m+2,\ldots,n\}$$

such that $1,2,\ldots,n$ occurs as a not-necessarily consecutive subsequence. Let $I_m(n)$ denote the cardinality of $\mathcal{I}_m(n)$. Then

$$I_{m+1}(n)=(n+2m)I_m(n)-m(n+m)I_{m-1}(n), \tag{1}$$

where $m\geq 1$, $n\geq 1$, $I_0(n)=1$, and $I_1(n)=n$.

H. S. Wilf pointed out that this is implied by a theorem of Gessel [G, p. 261].

Notation and definitions. We say that a permutation of a multiset has an *n-increasing subsequence*, if $1,2,\ldots,n$ is a not-necessarily consecutive subsequence of the permutation. We use $\uplus$ to denote multiset union. We call a letter x in a permutation of a multiset *inessential* if it appears in no n-increasing subsequence, otherwise *essential*. We use $[n]$ to denote the set $\{1,2,\ldots,n\}$. Let $x\in[n]$ and let S be a $(k-1)$-subset of $[n]$ that does not contain x. We define $\Phi(x,S)$ (respectively, $\Psi(x,S)$) to be the set of all permutations of the multiset $[n]\uplus S\uplus\{x\}$ with an n-increasing subsequence such that both copies of x are *essential* and the set of all letters that occur between the two copies of x is a *nonempty* (respectively, *possibly empty*) subset of S.

To prove the theorem, we need the following lemma.

LEMMA 1. *Let S be a $(k-1)$-subset of $[n]$ and let $\mathcal{T}$ be the set of permutations of $[n]\uplus S$ with an n-increasing subsequence. Then $|\mathcal{T}|=I_{k-1}(n)$.*

Proof. The result is true for all such sets S if it is true for one such. If $S=[k-1]$, however, then the assertion is just the definition of $I_{k-1}(n)$. □

2. Description of the sets $\mathcal{P}_{m+1}$ and $\mathcal{Q}_m$.

LEMMA 2. *Let $x\in[n]$ and let S be a $(k-1)$-subset of $[n]$ that does not contain x. Then $(n+k-1)I_{k-1}(n)-I_k(n)$ is the cardinality of $\Phi(x,S)$.*

Proof. Let $\mathcal{T}'$ (respectively, $\mathcal{T}''$) be the set of permutations of $[n]\uplus S$ (respectively, $[n]\uplus S\uplus\{x\}$) that have an n-increasing subsequence. We identify a set of permutations that are counted by the expression $(n+k-1)I_{k-1}(n)-I_k(n)$. Note that, by Lemma 1, $|\mathcal{T}'|=I_{k-1}(n)$ and $|\mathcal{T}''|=I_k(n)$. For each permutation σ' in $\mathcal{T}'$, we construct $n+k-1$ distinct permutations by successively inserting x into each of the $n+k$ available slots and noting that two of these slots, namely, the two that surround the copy of x that is already

*Received by the editors June 1, 1992; accepted for publication (in revised form) October 28, 1992. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

†Department of Mathematics, University of Pennsylvania, Philadelphia, Pennsylvania 19104–6395 (lily@rademacher.math.upenn.edu).

present, give the same result. As σ' runs through all elements of $\mathcal{T}'$, we create a multiset that is counted by $(n+k-1)I_{k-1}(n)$, since $|\mathcal{T}'| = I_{k-1}(n)$ by Lemma 1. It remains to identify the permutations that have been doubly created in this construction.

Let $\mathcal{F}_k$ be the set of $\tau \in \mathcal{T}''$ such that one of the two copies of $x \in \tau$ is *inessential.* We claim that a permutation τ_k has been double-counted if and only if $\tau_k \notin \mathcal{F}_k$ and its two copies of x are not adjacent. First, if $\tau_k \in \mathcal{F}_k$, then τ_k has been counted only once because the deletion of the x that is *inessential* yields a $\sigma' \in \mathcal{T}'$. The map is reversible. Second, suppose that $\tau_k \notin \mathcal{F}_k$ and that its two copies of x *are* adjacent. Then such a τ_k has been counted only once because the deletion of either copy of x yields the same τ. The map is also reversible. Third, if $\tau_k \notin \mathcal{F}_k$ and its two copies of x are not adjacent, then τ_k has been counted twice, which completes the proof of the claim. The permutations that are doubly counted are exactly all the permutations in $\Phi(x, S)$. Thus the lemma follows. □

Now we prove the theorem.

Proof of theorem. Let n be fixed, $\mathcal{I}_m(n)$ be denoted by $\mathcal{I}_m$, and $I_m(n)$ be denoted by I_m. We prove (1) or the equivalent form

$$(2) \qquad (n+m)I_m - I_{m+1} = m((n+m)I_{m-1} - I_m).$$

We identify a set, $\mathcal{P}_{m+1}$, of permutations that are counted by the left side of (2). By Lemma 2, with $S = [m]$ and $x = m+1$, we conclude that the left side of (2) is $|\Phi(m+1, [m])|$. Thus we can take $\mathcal{P}_{m+1} = \Phi(m+1, [m])$.

We identify a set $\mathcal{Q}_m$ of permutations that are counted by the right-hand side of (2). We claim that

$$\mathcal{Q}_m = \bigcup_{j=1}^{m} \Psi(j, [m] \setminus \{j\})$$

is such a set.

Proof of claim. First, we show that $|\Psi(x, S)| = (n+m)I_{m-1} - I_m$, where $x \in [m]$ and $S = [m] \setminus \{x\}$. Since $\Psi(x, S) \setminus \Phi(x, S)$ is the set of permutations of the multiset $[m] \uplus [n]$ with an n-increasing subsequence such that the two copies of x are adjacent and every permutation ρ in $\Psi(x, S) \setminus \Phi(x, S)$ corresponds to a permutation of the multiset $S \uplus [n]$ with an n-increasing subsequence simply by the deletion of a copy of x in ρ, $|\Psi(x, S) \setminus \Phi(x, S)| = I_{m-1}$. By Lemma 2, we obtain

$$\begin{aligned} |\Psi(x, S)| &= (n+m-1)I_{m-1} - I_m + I_{m-1} \\ &= (n+m)I_{m-1} - I_m. \end{aligned}$$

It remains to account for the factor of m on the right of (2). Since for each $x \in [m]$ and $S = [m] \setminus \{x\}$, we have $(n+m)I_{m-1} - I_m = |\Psi(x, S)|$. We thus conclude

$$\left| \bigcup_{x=1}^{m} \Psi(x, [m] \setminus \{x\}) \right| = m((n+m)I_{m-1} - I_m)$$

and

$$\mathcal{Q}_m = \bigcup_{x=1}^{m} \Psi(x, [m] \setminus \{x\}). \qquad □$$

3. Description of a bijection between $\mathcal{P}_{m+1}$ and $\mathcal{Q}_m$. Given a permutation where both copies of x are essential, we let the subscript F (respectively, L) denote the first (respectively, last) copy of x as x occurs in the permutation. We define a map $g : \mathcal{Q}_m \to \mathcal{P}_{m+1}$ as follows: Given a permutation in $\Psi(x, [m] \setminus \{x\})$, if $x < m$, g first inserts a new copy of $m+1$ into the position immediately preceding x_{L}, then it moves $m+1, x_{\mathrm{L}}$, and everything from x_{L} up to, but not including, the last essential copy of $x+1$, as a block, to the place immediately preceding the old copy of $m+1$. If $x=m$, then g simply inserts a new copy of $m+1$ into the place immediately before m_{L}.

To check that the image of g is in $\mathcal{P}_{m+1}$, we only need to check that both copies of $m+1$ are essential. First, we consider the case where $x<m$. The old copy of $m+1$ is essential because there is an n-increasing subsequence using x_{F}, the last essential copy of $x+1$, and the old $m+1$ before g is applied. Since g alters a block that is not used in the formation of an n-increasing subsequence and leaves the rest intact, the same n-increasing subsequence still exists after g is applied. The new copy of $m+1$ is essential because the same n-increasing subsequence using x_{F}, the last essential copy of $x+1$ in the old permutation, and the old $m+1$ can now use the new $m+1$ to form an n-increasing subsequence by the action of g. In the case where $x=m$, g does nothing but inserting a new copy of $m+1$ into the place immediately before m_{L}. Hence, the old copy of $m+1$ is still essential. The new copy of $m+1$ is also essential in an n-increasing subsequence using m_{F}.

We define a map $f:\mathcal{P}_{m+1}\to\mathcal{Q}_m$ as follows: (We will show that $f=g^{-1}$.) For every element in $\mathcal{P}_{m+1}$, f locates the (nonempty) segment sandwiched between the two copies of $m+1$ and notes the letter y that appears immediately after $(m+1)_{\mathrm{F}}$. If $y<m$, then f deletes $(m+1)_{\mathrm{F}}$ and moves the segment (excluding $m+1$'s) to the position immediately before the last essential copy of $y+1$ outside the segment. If $y=m$, then f simply deletes $(m+1)_{\mathrm{F}}$.

To check that the image of f is in $\mathcal{Q}_m$, we only need to show that both copies of y are essential. First, we consider the case where $y<m$. Before f is applied to a given permutation in $\mathcal{P}_{m+1}$, the permutation has an n-increasing subsequence such that $(m+1)_{\mathrm{L}}$ is used and, before $(m+1)_{\mathrm{F}}$, it contains y, $y+1$, and m, where the $y+1$ used is the last essential one before $(m+1)_{\mathrm{F}}$. Since f deletes $(m+1)_{\mathrm{F}}$ and moves the segment sandwiched between the $(m+1)$'s with y as its first member to the place immediately before the $y+1$ used in the n-increasing subsequence, the same n-increasing subsequence that does not contain $(m+1)_{\mathrm{F}}$ can use the newly allocated y instead of the old y. Thus both copies of y are essential. If $y=m$, then there is an essential m before $(m+1)_{\mathrm{F}}$, and an $m(=y)$ immediately after $(m+1)_{\mathrm{F}}$, before f is applied. After f is applied, i.e., the deletion of $(m+1)_{\mathrm{F}}$, there are two copies of m placed before $m+1$. Thus both copies of m are essential.

It is easy to check that $f\circ g$ is the identity mapping on $\mathcal{Q}_m$; $g\circ f$ is the identity mapping on $\mathcal{P}_{m+1}$. Thus, we have exhibited a bijection between $\mathcal{Q}_m$ and $\mathcal{P}_{m+1}$. □

An example of the actions of f and g. Let $n=9, m=5$, and

$$\rho=312324564567189.$$

The segment between two 6's is 45; thus $y=4$ and $y+1=5$. Since there is only one essential $y+1$ outside the segment between two 6's, f puts 45 in front of 5 and deletes 6 to obtain

$$\sigma=31232445567189.$$

Next, let $x=4$, $n=9$, $m=5$, and $S=[5]\setminus\{4\}=\{1,2,3,5\}$. Note that $x<m$. By the definition of g, 6 is inserted in front of the second copy of 4, then the block 645 is moved to the place immediately before 6, giving

$$g(\sigma)=312324564567189.$$

The example also shows that $g\circ f(\rho)=\rho$.

A generalization. I. M. Gessel found the following theorem.

THEOREM. *Let r be a positive integer. Then the number of permutations of the multiset*

$$\{1^{r+1}, 2^{r+1}, \ldots, m^{r+1}, (m+1)^r, \ldots, n^r\}$$

containing $1, 1, \ldots, 1, 2, 2, \ldots, 2, \ldots, n, n, \ldots, n$ *(r copies of every letter) as a not-necessarily consecutive subsequence is the coefficient of $x^m/m!$ in*

$$(1-x)^{-(rn+1)} \exp\left(\frac{-rx}{1-x}\right).$$

From Gessel's generating function, we have the following theorem.

THEOREM. *Let $\mathcal{I}_m^r(n)$ be the set of permutations of the multiset*

$$\{1^{r+1}, 2^{r+1}, \ldots, m^{r+1}, (m+1)^r, (m+2)^r, \ldots, n^r\}$$

such that $1, 1, \ldots, 1, 2, 2, \ldots, 2, \ldots, n, n, \ldots, n$ *(r copies of every letter) occurs as a not-necessarily consecutive subsequence. Let $I_m(n)$ denote the cardinality of $\mathcal{I}_m^r(n)$. Then*

$$I_{m+1}(n) = (rn + 2m - r + 1)I_m(n) - m(rn+m)I_{m-1}(n),$$

where $n \geq m \geq 1$, $r \geq 1$, $I_0(n) = 1$, and $I_1(n) = rn + 1 - r$.

In the following equivalent form of the recurrence

$$(rn + m + 1 - r)I_m - I_{m+1} = m((rn+m)I_{m-1} - I_m),$$

we can use the techniques in the proof of the theorem to identify two multisets and find a bijection between them.

Acknowledgments. I thank H. Wilf for introducing me to this problem and for a number of comments that have improved the presentation. I also thank I. M. Gessel for communicating his generalization to me.

REFERENCES

[G] I. M. GESSEL, *Symmetric functions and P-recursiveness*, J. Combin. Theory Ser. A, 53 (1990), pp. 257–285.

[GJ] I. P. GOULDEN AND D. M. JACKSON, *Combinatorial Enumeration*, John Wiley, New York, 1983.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 157–165, May 1994

INDIFFERENCE DIGRAPHS: A GENERALIZATION OF INDIFFERENCE GRAPHS AND SEMIORDERS*

M. SEN[†] AND B. K. SANYAL[‡]

Abstract. *Indifference digraphs*, an analogue of indifference graphs, are introduced. They are shown to be equivalent to two restricted classes of interval digraphs, *unit interval digraphs* and *proper interval digraphs*, also introduced in this paper. These digraphs are characterized in terms of their adjacency matrices and in terms of a generalized concept of a semiorder. The results generalize the corresponding results for undirected graphs and provide a generalization of the Scott–Suppes theorem characterizing semiorders.

Key words. intersection graph, indifference graph, interval graph, interval digraph

AMS subject classifications. 05C75, 06F99

1. Introduction. Indifference graphs were introduced and studied by Roberts [9]. An undirected graph is an *indifference graph* if there exists a real-valued function f on the vertices such that vertices u, v are adjacent if and only if $|f(u) - f(v)| \leq 1$. We may call f an *indifference representation* of G. Roberts characterized indifference graphs and proved that they are equivalent to *proper interval graphs* (intersection graphs of intervals in which no interval properly contains another) and to *unit interval graphs* (intersection graphs of unit-length intervals). In this paper, we generalize these results to directed graphs.

We first introduce an analogue of indifference graphs for directed graphs (henceforth "digraphs"). A digraph with edge set E is an *indifference digraph* if there exists an ordered pair of real-valued functions f, g on the vertices satisfying $uv \in E$ if and only if $|f(u) - g(v)| \leq 1$. Here $f(v)$ and $g(v)$ are called the *source value* and *sink value* of the vertex v, respectively, and the pair f, g is called an *indifference representation*. Interchanging the source value and sink value of each vertex shows that the converse of an indifference digraph (obtained by changing the direction of each edge) is also an indifference digraph. A complete digraph (that is, a digraph in which every ordered pair of vertices forms an edge including the loops) is also an indifference digraph, as can be seen by assigning arbitrary source and sink values all within the interval [0, 1].

Beineke and Zamfirescu [1] and Sen et al. [11] introduced and studied (in different contexts) a natural analogue of intersection graph models for directed graphs. Let V be a finite family of ordered pairs of sets and assign a vertex v to each ordered pair. The first set assigned to v is called its *source set* S_v, and the second set is its *sink set* T_v. The *intersection digraph* of the family of pairs is the digraph $D = (V, E)$ such that $uv \in E$ if and only if $S_u \cap T_v \neq \emptyset$. An *interval digraph* is the intersection digraph of a family of ordered pairs of intervals on the real line; these were characterized in [11] and [12].

By placing constraints on the source and sink intervals, we introduce two types of interval digraphs. *Unit interval digraphs* are interval digraphs with interval representations such that all the source intervals and sink intervals have unit length. *Proper interval digraphs* are interval digraphs with representations such that no source interval properly contains another source interval, and no sink interval properly contains another sink interval. Note that there is no restriction on the inclusion relationship between a source interval S_u and a sink interval T_v.

* Received by the editors January 31, 1990; accepted for publication (in revised form) August 20, 1992.

[†] Department of Mathematics, North Bengal University, Darjeeling, Pin 734 430, West Bengal, India.

[‡] Department of Mathematics, University College, Raiganj, W. Dinajpur, Pin 733 134, West Bengal, India.

Given an indifference representation f, g, setting $S_u = [f(u) - \frac{1}{2}, f(u) + \frac{1}{2}]$ and $T_u = [g(u) - \frac{1}{2}, g(u) + \frac{1}{2}]$ for all u provides an interval representation. Hence every indifference digraph is an interval digraph; indeed, it is a unit interval digraph, since these intervals have length 1. Conversely, setting $f(u)$, $g(u)$ to the midpoints of S_u, T_u in a unit interval representation yields an indifference representation, so indifference digraphs are precisely the unit interval digraphs. Section 2 contains an explicit interval digraph that is not an indifference digraph.

Making use of the characterizations of interval digraphs in [11], we give characterizations of the more restricted classes to prove that the classes of indifference digraphs, unit interval digraphs, and proper interval digraphs are all the same. The most important characterization, from which others are obtained, is a characterization of the adjacency matrices of these digraphs. We say that a 0,1-matrix has a *monotone consecutive arrangement* if there exist independent row and column permutations exhibiting the following structure: The 0's of the resulting matrix can be labeled R or C such that every position above or to the right of an R is an R, and every position below or to the left of a C is a C. The name arises from an equivalent restatement of the condition described in §2.

Roberts [9] proved equivalence for the analogous classes of undirected graphs. Our results reduce to the earlier results on undirected graphs when we view undirected graphs as symmetric digraphs with loops (i.e., symmetric and reflexive binary relations). The adjacency matrix of the corresponding digraph is obtained by adding 1's on the diagonal; this is called the *augmented adjacency matrix* $A^*(G)$ for an undirected graph G. A symmetric digraph with loops has an indifference representation if and only if it has an indifference representation with $f = g$, because the symmetry implies that averaging f and g will not change the resulting edges. Conversely, every indifference representation with $f = g$ yields a symmetric digraph with loops. This establishes a bijection between indifference graphs and indifference digraphs representable using $f = g$.

Hence our characterization implies that G is an indifference graph if and only if $A^*(G)$ has a monotone consecutive arrangement. This reduces to the Roberts characterization [8] that G is an indifference graph if and only if $A^*(G)$ has a column permutation so that the 1's appear consecutively in each row (called the *consecutive ones property for rows*). A monotone consecutive arrangement exhibits the consecutive ones property for both rows and columns. Conversely, the symmetry of A^* allows us to apply any column permutation also to the rows to achieve the consecutive ones property for each simultaneously, while leaving 1's on the diagonal; this is a monotone consecutive arrangement.

In fact, this correspondence between indifference graphs and special indifference digraphs holds in the more general setting of interval graphs. A graph G is an interval graph if and only if the corresponding symmetric digraph with loops G^* is an interval digraph, because an interval representation for G^* becomes an interval representation for G when $S_v = [a, b]$ and $T_v = [c, d]$ are replaced by $I_v = [(a + c)/2, (b + d)/2]$. The details of this proof will appear in [13]. The effect is that some of the results about interval graphs are actually special cases of the results about interval digraphs in [11] and [12].

Roberts introduced indifference graphs as a graph-theoretic concept related to semiorders, which are a special type of binary relation. In discussing a binary relation P, we use the notation xPy, $xy \in P$, and $x \to y$ interchangeably, corresponding to the following several equivalent notions: (1) P as a binary relation, (2) P as the set of ordered pairs of a relation, and (3) P as the edges of a digraph. We use whichever notation is convenient. Luce [6] and Scott and Suppes [10] defined a *semiorder* to be an irreflexive binary relation (loopless digraph) satisfying (1) ab, $cd \in P$ implies aPd or cPb and (2) ab,

$bc \in P$ implies aPd or dPc, where, in each case, a, b, c, d are arbitrary (not necessarily distinct) elements (or vertices). The Scott–Suppes theorem [10] characterizes semiorders as those binary relations P for which there exists a real-valued function f such that xPy if and only if $f(x) > f(y) + \delta$ for some constant δ (which can be taken to be 1). This condition expresses P as a transitive orientation of the complement of an indifference graph; hence we call f a *coindifference representation.* When viewing an indifference graph as a symmetric digraph with loops, this result was rephrased by Roberts [9] by saying that a graph with edges E is an indifference graph if and only if there is a semiorder P such that $\bar{E} = P \cup P^{-1}$, where P^{-1} is the digraph obtained from P by reversing the directions on all the edges.

We introduce a generalization of semiorders that behaves in the analogous way for indifference digraphs. We obtain the Scott–Suppes theorem and Roberts's rephrasing of it as special cases of the resulting theorem. We define a *generalized semiorder* to be a pair of disjoint binary relations (edge-disjoint digraphs) H_1, H_2 on the same set such that (1) aH_ib and cH_id imply aH_id or cH_ib ($i \in \{1, 2\}$), (2) aH_1b and cH_2b imply aH_1d or cH_2d, and (3) aH_1b and aH_2c imply dH_1b or dH_2c, where a, b, c, d are arbitrary (not necessarily distinct) elements.

Observe that if $H_2 = H_1^{-1}$, then conditions (2) and (3), directly above, coincide, and disjointness forbids loops, so that H_1, H_2 are semiorders. We prove that a digraph with edge set E is an indifference digraph if and only if there is a generalized semiorder H_1, H_2 on the vertex set such that $\bar{E} = H_1 \cup H_2$. In particular, we prove the following generalization of the Scott–Suppes theorem: (H_1, H_2) is a generalized semiorder on the elements A if and only if there are two functions f, g: $A \rightarrow \mathbf{R}$ such that $xH_1y \Leftrightarrow f(x) > g(y) + 1$ and $xH_2y \Leftrightarrow g(y) > f(x) + 1$. The two functions f, g are called a *coindifference representation.*

To obtain the Scott–Suppes theorem as a corollary, suppose that H is a binary relation and let $H_1 = H$, $H_2 = H^{-1}$. If H has a coindifference representation f (with threshold 1), then xH_1y if and only if $f(x) > f(y) + 1$ and similarly xH_2y if and only if $f(y) > f(x) + 1$. Hence (H_1, H_2) is a generalized semiorder, which implies that H is a semiorder as noted above. Conversely, suppose that H is a semiorder. This implies that (H_1, H_2) is a generalized semiorder, so we obtain a coindifference representation f, g satisfying $xH_1y \Leftrightarrow f(x) > g(y) + 1$ and $yH_2x \Leftrightarrow g(x) > f(y) + 1$. Let $h(x) = [f(x) + g(x)]/2$. If xHy, then $h(x) > h(y) + 1$. If $h(x) > h(y) + 1$, then xH_1y or yH_2x, i.e., xHy, and h is a coindifference representation of H.

We note that other generalizations of semiorders have been introduced, including the *bisemiorder* of Ducamp and Falmagne [4] and the *double semiorder* of Cozzens and Roberts [3]. In both cases, some analogues of the results on undirected graphs were obtained.

2. Properties of adjacency matrices. We observed earlier that every indifference digraph is an interval digraph. We will need the characterizations of interval digraphs. First, a digraph is a *Ferrers digraph* or *Ferrers relation* if it satisfies condition (1) in the definition of semiorder: ab, $cd \in P$ implies aPd or cPb. This definition, introduced by Riguet [7], is equivalent to forbidding the adjacency matrix to have a 2×2 submatrix that is a permutation matrix, or to requiring the successor sets (or predecessor sets) to be ordered by inclusion. The name arises from another characterization: The rows and columns of the adjacency matrix of a Ferrers digraph can be independently permuted so that the positions of the 1's form a Ferrers diagram.

THEOREM 1 (see [11], [12]). *For a digraph D, the following are equivalent:*

(1) *D is an interval digraph;*

(2) *$\bar{D}$ is the union of two disjoint Ferrers digraphs;*

(3) *The rows and columns of the adjacency matrix of D can be permuted independently such that each* 0 *can be labeled with R or C in such a way that every position to the right of an R is an R and every position below a C is a C.*

A matrix satisfying condition (3), above, has the *partitionable zeros property*; note that it is a weaker form of the monotone consecutive arrangement condition described earlier. This characterization enables us to show that the indifference digraphs are properly contained in the interval digraphs.

Example 1 (an interval digraph that is not an indifference digraph). The adjacency matrix below satisfies the characterizations above, so the corresponding digraph is an interval digraph:

$$\begin{matrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}.$$

Let v_1, v_2, v_3, v_4 denote the vertices in order and suppose that there is an indifference representation with $f(v_i) = f_i$ and $g(v_i) = g_i$. All sink values must lie in the interval $[f_2 - 1, f_2 + 1]$. Since edges are determined by absolute value of $f_i - g_j$, we may assume that $g_4 \geq f_2$. Now the 0's in rows 1 and 3 force $f_1, f_3 < f_2$, and then $g_3, g_1 > f_2$. However, a 2×2 permutation submatrix has no indifference representation in which both source values are less than both sink values; the larger source value must be more than 1 away from one of the sink values, which then cannot be within distance 1 of the smaller source value.

A more straightforward characterization of interval digraphs in [11] can be restricted in a natural way to characterize proper interval digraphs. A *generalized complete bipartite subdigraph* (GBS) of a digraph is a subdigraph consisting of vertex sets X, Y and edges from all of X to all of Y. The word "generalized" indicates the fact that X, Y need not be disjoint; any submatrix of 1's in the adjacency matrix yields a GBS. If $\mathbf{B} = \{(X_k, Y_k)\}$ is a collection of GBSs, we can form the incidence matrix between the vertices V and the source sets $\{X_k\}$, called the *V, X-matrix*, and similarly the *V, Y-matrix* between the vertices and sink sets. Since the rows in the two matrices can be viewed as source sets and sink sets, with the GBSs corresponding to elements that can be placed in order along a line, we obtain the following result.

THEOREM 2 (see [11]). *A digraph is an interval digraph if and only if its edges can be covered by a collection of* GBS*s* **B** *that can be indexed so that the* 1*'s in each row of the V, X-matrix and the V, Y-matrix appear consecutively.*

To obtain proper sets of intervals, we want the resulting matrices to have the *proper consecutive ones property* (for rows) introduced by Fishburn [5]. This is defined to be the existence of a column ordering so that the 1's in each row appear consecutively and do not properly contain the 1's in any other row. Such an ordering "exhibits" the property. By transforming a proper interval representation into GBSs corresponding to the endpoints of all intervals, and conversely by transforming membership in an appropriate sequence of GBSs into intervals, we obtain the following result.

THEOREM 3. *A digraph is a proper interval digraph if and only if its edges can be covered by a collection of* GBS*s* **B** *that can be indexed so that the* 1*'s in the rows of the V, X-matrix and in the rows of the V, Y-matrix exhibit the proper consecutive ones property.*

Note that, given a matrix with the proper consecutive ones property for rows, we can permute the rows to exhibit a monotone consecutive arrangement. The converse does not hold, as a monotone consecutive arrangement allows proper inclusion of 1's in rows. We see in Theorem 5 that the condition of Theorem 3 for proper interval digraphs is equivalent to requiring the weaker monotone consecutive condition for the incidence matrices of some family of *maximal* GBSs that cover D. We will need a rephrasing of the monotone consecutive condition that focuses explicitly on how the 1's in rows must behave. The following characterization is the source of the name "monotone consecutive arrangement." Note that zero rows or columns can be placed at the bottom or right without affecting the existence of a monotone consecutive arrangement.

LEMMA 1. *A* 0,1-*matrix with n nonzero rows has a monotone consecutive arrangement if and only if it has independent row and column permutations such that the* 1*'s appear consecutively in each row and the values* $\{a_i\}$ *and* $\{b_i\}$ *denoting the initial column and final column of the interval of* 1*'s in row i satisfy* $a_1 \leq \cdots \leq a_n$ *and* $b_1 \leq \cdots \leq b_n$.

Proof. The labeling of the 0's in a monotone consecutive arrangement guarantees $a_{i-1} \leq a_i$ and $b_{i+1} \geq b_i$. Conversely, if M is permuted so sequences a and b are monotone, let a 0 in position (i, j) have label C if $j < a_i$, or label R if $j > b_i$. By construction, the labeling condition is satisfied in rows, and monotonicity of the sequences guarantees that it is satisfied in columns. □

Finally, it is worthwhile to note that, although a monotone consecutive arrangement implies the consecutive ones property for both rows and columns, the converse does not hold.

Example 1 (continued). The adjacency matrix shown in the previous portion of this example has the consecutive ones property for both rows and columns, but it does not yield an indifference digraph. By the characterization we claim, it thus does not have a monotone consecutive arrangement. For clarity and motivation, we give a short direct proof of that. In a monotone consecutive arrangement, the 0's of a 2×2 permutation submatrix cannot be both C or both R. By symmetry of v_1 and v_3 in this digraph, we may assume that the 0 in position (1, 3) is labeled R and that the 0 in position (3, 1) is labeled C. Since row 2 has all 1's, it thus must be under row 1 and over row 3 in any monotone consecutive arrangement. This implies that the 0's of column 4 cannot be both C or both R. With one of each in these rows, column 4 must occur after column 2 and before column 2 in any monotone consecutive arrangement, which is impossible.

3. Equivalence of digraph classes. In this section, we prove characterizations of the digraph classes we have been considering. First, we characterize the adjacency matrices and show that the classes are equivalent.

THEOREM 4. *If D is a digraph, the following conditions are equivalent*:

(a) *D is an indifference digraph*;

(b) *D is a unit interval digraph*;

(c) *D is a proper interval digraph*;

(d) *The adjacency matrix of D has a monotone consecutive arrangement.*

Proof. We noted (a) ⇔ (b) ⇒ (c) in the Introduction. For (c) ⇒ (d), suppose that D is a proper interval digraph. By Theorem 3, we have a collection $\mathbf{B} = \{(X_k, Y_k)\}$ of GBSs that cover D, such that, for the V, X- and V, Y-matrices, the 1's in each row appear consecutively and do not properly contain the 1's of any other row. We may assume that each row has a 1; otherwise, the corresponding row or column of the adjacency matrix can be placed at the end, which will not affect the existence of a monotone arrangement.

Let $a(v)$, $b(v)$, respectively, denote the first and last columns containing 1 in the row of the V, X-matrix corresponding to v; similarly, define $c(v)$, $d(v)$ from the V, Y-matrix. Index the vertices $u_1, \ldots, u_n$ so that $a(u_1) \leq \cdots \leq a(u_n)$; the proper consecutive ones property also implies that $b(u_1) \leq \cdots \leq b(u_n)$. Similarly, index them as $v_1, \ldots, v_n$ so that $c(v_1) \leq \cdots \leq c(v_n)$ and $d(v_1) \leq \cdots \leq d(v_n)$. Let $a_i = a(u_i)$, $b_i = b(u_i)$, $c_j = c(v_j)$, $d_j = d(v_j)$. We may assume that all source sets and sink sets are nonempty, which means that, for any $(X_k, Y_k) \in \mathbf{B}$, there exist i, j such that $k \in [a_i, b_i]$ and $k \in [c_j, d_j]$.

We claim that the row ordering $u_1, \ldots, u_n$ and column ordering $v_1, \ldots, v_n$ of the adjacency matrix exhibit a monotone consecutive arrangement. The edges with tail u_i are covered by the GBSs with source sets $X_{a_i}, \ldots, X_{b_i}$; hence the successors (out-neighbors) of u_i are $Y_{a_i} \cup \cdots \cup Y_{b_i}$. Let $\alpha_i = \min\{j\colon a_i \in [c_j, d_j]\}$ and $\beta_i = \max\{j\colon b_i \in [c_j, d_j]\}$. The proper consecutive ones property of the V, Y-matrix (and lack of 0 rows, except possibly at the end) implies that the ones in row i of the adjacency matrix are $\{j\colon \alpha_i \leq j \leq \beta_i\}$. It also implies that $\alpha_i \leq \alpha_{i+1}$ because $a_i \leq a_{i+1}$, and that $\beta_i \leq \beta_{i+1}$ because $b_i \leq b_{i+1}$. By Lemma 1, we have a monotone consecutive arrangement.

To prove (d) $\Rightarrow$ (a), we take a monotone consecutive arrangement of an m by n 0,1-matrix M with rows $u_1, \ldots, u_m$ and columns $v_1, \ldots, v_n$ and construct an indifference representation, writing $f_i = f(u_i)$ and $g_j = g(v_j)$. To avoid technical degeneracies, we use induction on $m + n$. In particular, if the matrix has a $\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}$ decomposition, then we can take representations of the two smaller matrices by induction and shift one by a large constant to obtain a representation of M (similarly for zero rows or columns). Hence we may assume that every row and column has a 1 (including position (1, 1)) and that each consecutive pair of rows have 1's in some common column.

From the monotone consecutive arrangement, we have nondecreasing sequences $\{a_i\}$ and $\{b_i\}$ such that the successor set of u_i is $Q_i = \{v_{a_i}, \ldots, v_{b_i}\}$; the existence of a common 1 is equivalent to $a_i \leq b_{i-1}$. We distinguish the *maximal* successor sets, those not contained in any other. We present an iterative algorithm such that, at the end of stage i, we have specified $g_1 < \cdots < g_{b_i}$ and $f_1 \leq \cdots \leq f_i$ to form an indifference representation of the first i rows and g_{b_i} columns. Furthermore, the values are chosen so that $g_{b_i} - 1 = f_i = g_{a_i} + 1$ if Q_i is maximal and $g_{b_i} - 1 < f_i < g_{a_i} + 1$ if Q_i is not maximal. As an initialization, we can specify any desired value; we choose $g_1 = 1$ and by convention we take $b_0 = 1$ and $g_0 = -1$ to treat stage 1 inductively.

To begin stage i, we set the sink values through g_{b_i}. If $b_i = b_{i-1}$, then these are already known. If $b_i > b_{i-1}$, then we need to set the sink values above b_{i-1}. Since $a_i \leq b_{i-1}$, the value g_{a_i} has been set. Choose $g_{b_{i-1}+1}, \ldots, g_{b_i}$ as an arbitrary increasing sequence in the open interval $(f_{i-1} + 1, g_{a_i} + 2)$, except set $g_{b_i} = g_{a_i} + 2$ if Q_i is maximal. This interval is nonempty because $f_{i-1} \leq a_{i-1} + 1 \leq a_i + 1$, and the first inequality is strict if Q_{i-1} is not maximal, while the second is strict if Q_{i-1} is maximal and $b_i > b_{i-1}$. The resulting inequality $g_{b_{i-1}+1} > f_{i-1} + 1$ guarantees that the representation is correct on the new columns and old rows. Note also that $f_{i-1} + 1 \geq g_{b_{i-1}}$, so the sink values remain increasing.

To complete stage i, we choose f_i to add row i to the representation and preserve the stated restrictions. We must have $g_{b_i} - 1 \leq f_i \leq g_{a_i} + 1$ (with strict inequalities if Q_i is not maximal and equalities if Q_i is maximal), and $f_i > g_{a_i - 1} + 1$. Let $f_i = (\lambda + \mu)/2$, where $\lambda = \max\{g_{a_i - 1} + 1, g_{b_i} - 1\}$ and $\mu = g_{a_i} + 1$. By the construction of g_{b_i}, we have $\lambda \leq \mu$, with equality if and only if Q_i is maximal. This guarantees that f_i satisfies all the restrictions, and the monotonicity of $\{a_i\}$ and $\{b_i\}$ guarantees $f_i \geq f_{i-1}$. □

Example 2. Below is a monotone consecutive arrangement and an indifference representation that is an outcome of the algorithm described in the proof:

$$\begin{array}{cccccccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0\text{’} \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}$$

i	1	2	3	4	5	6	7	8	9	10	11	12
$f(u_i)$	1	2	3.8	3.9	4.5	4.9	5	5.9	6.3	6.8	7	7.9 ·
$g(v_i)$	1	2.6	2.8	3	4	4.8	5	5.6	5.8	6	7.8	8

Using the adjacency matrix characterization, we can now give an alternate characterization of proper interval digraphs as a variation on Theorem 3. We can restrict our attention to maximal GBSs if we allow the flexibility of a monotone consecutive arrangement of the incidence matrices instead of requiring the proper consecutive ones property.

THEOREM 5. *A digraph D is a proper interval digraph* (*or indifference digraph*) *if and only if it has a covering by a family of maximal* GBSs $\mathbf{B} = \{(X_k, Y_k)\}$ *numbered so that the V, X-matrix and V, Y-matrix exhibit monotone consecutive arrangements.*

Proof. For necessity, we may assume that we have a monotone consecutive arrangement of the adjacency matrix, with $[a_i, b_i]$ being the interval of columns with 1's in row i. Again, we implicitly use induction on $m + n$ to avoid degeneracies and assume that there is a 1 in every row and column and a common column in any two consecutive rows. There is a natural set of maximal GBSs associated with such a configuration. As suggested by viewing the matrix in the example, these are the maximal rectangular blocks of ones determined by corners of the region of 1's in the matrix. The upper corners are the positions of the form (i, b_i) with $b_i > b_{i-1}$; the lower corners have the form (j, a_j) with $a_j < a_{j+1}$. If $i \leq j$ and $a_j \leq b_i$, then these form a maximal GBS $B(i, j)$.

To select an appropriate family $\{B_k\}$, begin with $B_1 = B(1, r)$, where $r = \max\ \{i\colon a_i = 1\}$. Having determined $B_{k-1} = B(c_{k-1}, d_{k-1})$, let α be the next row below c_{k-1} having an upper corner (if any) and let β be the next row below d_{k-1} having a lower corner (if any). Since we have avoided degeneracy, we must have $\alpha \leq d_{k-1}$ or $a_\beta \leq b_{c_{k-1}}$ (unless B_{k-1} completes the covering), which implies that $B(\alpha, d_{k-1})$ or $B(c_{k-1}, \beta)$ can be chosen as $B_k = B(c_k, d_k)$. Since we shift by one corner at a time, the resulting sequence covers all the 1's.

In this sequence, the vertices of X_k are the consecutive set $u_{c_k}, \ldots, u_{d_k}$. Since each step we take increases c_k or d_k while leaving the other fixed, $\{c_k\}$ and $\{d_k\}$ are monotone, and the transpose of the V, X-matrix has a monotone consecutive arrangement. Since the original definition is invariant under transpose, the V, X-matrix also has a monotone consecutive arrangement. Applying the same argument to the transpose of the original adjacency matrix shows that the V, Y-matrix of $\mathbf{B}$ also has a monotone consecutive arrangement.

For sufficiency, let **B** be a collection of GBSs covering D that is indexed so the V, X-matrix and V, Y-matrix exhibit monotone consecutive arrangements. By Theorem 3, it suffices to show that, if there is any proper inclusion between the 1's of a consecutive pair of rows in the V, X-matrix or V, Y-matrix, then we can add one GBS to the sequence to reduce the number of proper inclusions.

By symmetry, we may assume that there is such an inclusion in the V, X-matrix, with $a_i < a_{i+1}$ but $b_i = b_{i+1} = k$. Define a new GBS (X', Y') by $X' = X_k - \{u_j: j \leq i\}$ and $Y' = Y_k$ and insert this into the sequence immediately following (X_k, Y_k). Now $b_i < b_{i+1}$, and the other such inequalities remain unchanged. □

4. Generalized semiorders and indifference digraphs. Using Theorem 4, it is easy to prove the generalization of the Roberts restatement of the Scott–Suppes theorem, as described in the Introduction. The generalization of the Scott–Suppes theorem itself follows as a corollary. Complementation is taken with respect to the complete relation $V \times V$, including loops.

THEOREM 6. *A digraph $D = (V, E)$ is an indifference digraph if and only if $\bar{E} = H_1 \cup H_2$, where (H_1, H_2) is a generalized semiorder on V.*

Proof. Necessity follows easily from Theorem 4 by considering a monotone consecutive arrangement of the adjacency matrix. Let $[a_i, b_i]$ be the interval of 1's in row i. Let H_1 correspond to the 0's labeled R (those after b_i) and H_2 to the 0's labeled C (those before a_i). Hence $\bar{E} = H_1 \cup H_2$, and H_1, H_2 are disjoint. Monotonicity implies that H_1 and H_2 are Ferrers digraphs, which is equivalent to the generalized semiorder condition (1) that uH_iv and xH_iy imply uH_iy or xH_iv ($i \in \{1, 2\}$).

For the other conditions, let u_i, u_k be the (not necessarily distinct) source vertices and let v_j, v_l be the (not necessarily distinct) sink vertices, indexed as in the monotone consecutive arrangement. To prove that (2) $u_iH_1v_j$ and $u_kH_2v_j$ imply $u_iH_1v_l$ or $u_kH_2v_l$, note that the hypothesis implies $j > b_i$ and $j < a_k$. If $l \geq j$, then $l > b_i$ and $u_iH_1v_l$, while, if $l \leq j$, then $l < a_k$ and $u_kH_2v_l$. When we appeal to the fact that the transpose of a monotone consecutive arrangement is also a monotone consecutive arrangement, the proof is the same for (3) $u_iH_1v_j$ and $u_iH_2v_l$ imply $u_kH_1v_j$ or $u_kH_2v_l$.

For sufficiency, suppose that $\bar{E} = H_1 \cup H_2$, where (H_1, H_2) is a generalized semiorder on V. As noted in the Introduction, we can view H_1, H_2 as relations or as digraphs on V; let $d_i^+(u)$ denote the set of successors (out-degree) of u in H_1; similarly, define d_i^- for in-degree. We claim it is not possible to have both $d_1^+(u) > d_1^+(u')$ and $d_2^+(u) > d_2^+(u')$. If these hold, then there are vertices v, v' such that $uv \in H_1$ but $u'v \notin H_1$ and $uv' \in H_2$ but $u'v' \notin H_2$, but this is precisely the configuration forbidden by condition (3). Similarly, the combination $d_1^-(v) > d_1^-(v')$ and $d_2^-(v) > d_2^-(v')$ is forbidden by condition (2). This implies that we can order the elements as $u_1, \ldots, u_n$ and as $v_1, \ldots, v_n$ so that, for $1 \leq i < n$, we have $d_1^+(u_i) \leq d_1^+(u_{i+1})$, $d_2^+(u_i) \geq d_2^+(u_{i+1})$, $d_1^-(v_i) \geq d_1^-(v_{i+1})$, and $d_2^-(v_i) \leq d_2^-(v_{i+1})$. By condition (1), H_1 and H_2 are Ferrers digraphs, and this ordering of the degrees simultaneously places the relations of H_1 in the upper right and H_2 in the lower left so that any position above or to the right of a position in H_1 is also in H_1, and any position above or to the right of a position in H_2 is also in H_2. In other words, this ordering of the rows and columns is a monotone consecutive arrangement. □

The generalization of the Scott–Suppes theorem is just a rephrasing of this result.

COROLLARY 1. *A pair (H_1, H_2) of relations on a set is a generalized semiorder if and only if it has a coindifference representation f, g; that is, real-valued functions f, g exist on the elements so that xH_1y if and only if $f(x) > g(y) + 1$, and xH_2y if and only if $g(y) > f(x) + 1$.*

Acknowledgment. We thank Douglas B. West for suggestions about presentation and organization of this material and for writing the current version of the paper.

REFERENCES

[1] L. W. BEINEKE AND C. M. ZAMFIRESCU, *Connection digraphs and second order line graphs*, Discrete Math., 39 (1982), pp. 237–254.

[2] M. B. COZZENS, *Higher and Multidimensional Analogues of Interval Graphs*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1981.

[3] M. B. COZZENS AND F. S. ROBERTS, *Double semiorders and double indifference graphs*, SIAM J. Algebraic Disc. Meth., 3 (1982), pp. 566–582.

[4] A. DUCAMP AND J. C. FALMAGNE, *Composite measurement*, J. Math. Psych., 6 (1969), pp. 359–390.

[5] P. C. FISHBURN, *Interval Orders and Interval Graphs*, John Wiley, New York, 1985.

[6] R. D. LUCE, *Semiorders and a theory of utility discrimination*, Econometrica, 24 (1956), pp. 178–191.

[7] J. RIGUET, *Les relations des Ferrers*, C.R. Acad. Sci. Paris, 232 (1951), p. 1729.

[8] F. S. ROBERTS, *Representations of Indifference Relations*, Ph.D. thesis, Stanford University, Stanford, CA, 1968.

[9] ———, *Indifference graphs*, in Proof Techniques in Graph Theory, F. Harary, ed., Academic Press, New York, 1969, pp. 139–146.

[10] D. SCOTT AND P. SUPPES, *Foundational aspects of theories of measurement*, J. Symbolic Logic, 23 (1958), pp. 233–247.

[11] M. SEN, S. DAS, A. B. ROY, AND D. B. WEST, *Interval digraphs—An analogue of interval graphs*, J. Graph Theory, 13 (1989), pp. 189–202.

[12] M. SEN, S. DAS, AND D. B. WEST, *Circular-arc digraphs*, J. Graph Theory, 13 (1989), pp. 581–592.

[13] M. SEN, B. K. SANYAL, AND D. B. WEST, *Representations of digraphs using intervals and circular arcs*, submitted.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 166–171, May 1994

SOME NEW BOUNDS ON SINGLE-CHANGE COVERING DESIGNS*

GUO-HUI ZHANG†

Abstract. A single-change covering design SCD (v, k, b) on a v-set V is a sequence of b k-sets (blocks) on V that together cover every pair of elements of V at least once, such that two successive blocks have $k - 1$ elements in common. Let $f(v, k)$ denote the smallest b for which there exists an SCD (v, k, b). Some new upper and lower bounds for $f(v, k)$ are given.

Key words. covering designs, design with a hole, single-change

AMS subject classification. 05

1. Introduction. Given positive integers v, k, and λ, a (v, k, λ)-*covering design* is a collection of k-subsets (blocks) of a given v-set V in which every pair of elements of V occur together in at least λ of the blocks. For a general discussion of covering designs, see, for example, [2]. An *ordered* (v, k, λ)-*covering design* is a (v, k, λ)-covering design, together with a way to order the blocks of the design. A *single-change ordered block design* SBD (V, B), or SBD (v, k, b), on a v-set V is a sequence B of b k-sets (blocks) on V such that two successive blocks have $k - 1$ elements in common. A *single-change covering design* SCD (V, B) (or SCD (v, k, b)) is an SBD (V, B) (or SBD (v, k, b)) such that each pair of elements of V occur together in at least one block. An SCD (v, k, b) is called *perfect* if $b' \geqq b$ for every SCD (v, k, b'). We will use $f(v, k)$ to denote the number of blocks in a perfect SCD (v, k, b).

Ordered covering designs arise in the testing of electrical components for compatibility. For a more detailed explanation of their applications, refer to [4], where it was noted that economical single-change covering designs provide an ideal model for such a testing. The following two results were proved in [4].

LEMMA 1.1 (see [4]). $f(v, k) \geqq \lceil (v(v-1) - (k-1)(k-2))/2(k-1) \rceil$, *with equality if* $k = 2$ *or* 3.

LEMMA 1.2 (see [4]). $f(v, k) \geqq 2(v - k) + 1$, *with equality if* $v \leqq 2k$ *and* $k \geqq 3$.

A *single-change covering design with a hole* SCDH (V, B), or SCDH (v, k, b), is an SBD (V, B) (or SBD (v, k, b)) such that every pair of elements of V occur together in at least one of the blocks, except for one pair called a *hole*, say ab, which has the property that one of a and b appear in either the first or the last block of B. We use $f'(v, k)$ to denote the smallest b for which there exists an SCDH (v, k, b). Trivially, $f(v, k) \leqq f'(v, k) + 1$. Hence any upper bound for $f'(v, k)$ will also provide one for $f(v, k)$.

It should be mentioned that the idea of using designs with "holes" to study original designs is typical in combinatorial designs. For example, orthogonal Latin squares with "holes," called *incomplete orthogonal arrays*, were studied in [1]; Kirkman triple systems with "holes" were referred to as *frames* and studied in [3].

The proof of the following result is parallel to that of Lemma 1.1.

LEMMA 1.3. $f'(v, k) \geqq \lceil (v(v-1) - (k-1)(k-2) - 2)/2(k-1) \rceil$, *with equality if* $k = 2$ *or* 3.

In § 2 an upper bound for $f(v, k)$ is given by studying $f'(v, k)$. As a consequence, a tight bound for $f(v, k)$ is obtained in the case of $k = 4$. In § 3 a lower bound for $f(v, k)$ is given, which is perfect in many cases. Finally, we discuss a recursive method for estimating $f(v, k)$ in § 4, where a conjecture is also formulated.

* Received by the editors January 13, 1992; accepted for publication (in revised form) August 23, 1992.

† Department of Mathematical Sciences, University of Alabama at Huntsville, Huntsville, Alabama 35899.

2. An upper bound. Assume that $v = n(k-1) + i$, where $k+1 \leqq i \leqq 2k-1$, $n \geqq 0$, and $k \geqq 4$. Let $d(v, k) = 2(i-k) + n(i+k-3) + n(n-1)(k-1)/2$. Then we have the following result.

THEOREM 2.1. *It holds that* $f'(v, k) \leqq d(v, k)$.

Proof. We use induction on n.

Case 1. $n = 0$. Let the $v\,(=i)$ treatments be $1, 2, \ldots, i$. Then the following sequence S_0 of blocks shows that $f'(v, k) \leqq 2(v-k) = d(v, k)$, where $1i$ is the hole:

$$\begin{array}{l}
(i-k+1)\cdots(k-1)k\cdots i\\
\quad(i-k)\cdots(k-1)(k+1)\cdots i\\
\qquad\qquad\vdots\\
\qquad\qquad 3\cdots(k-1)(i-2)(i-1)i\\
\qquad\qquad 2\cdots(k-1)(i-1)i\\
\qquad\qquad 1\cdots(k-1)(i-1)\\
\qquad\qquad 1\cdots(k-1)(i-2)\\
\qquad\qquad\vdots\\
\qquad\qquad 1\cdots(k-1)k.
\end{array}$$

Case 2. $n = 1$. Let the v treatments be $1, \ldots, i, a_1, \ldots, a_{k-1}$. The blocks in S_0 of Case 1 plus the following sequence S_1 of blocks shows that $f'(v, k) \leqq d(v, k)$ with $3a_{k-1}$ as a hole:

$$\begin{array}{l}
\qquad 1\cdots(k-1)a_1\\
\qquad 1\cdots(k-1)a_2\\
\qquad 1\cdots(k-2)a_2a_3\\
\qquad 1\cdots(k-3)a_2a_3a_4\\
\qquad\qquad\vdots\\
\quad 1\ 2\ a_2\cdots a_{k-1}\\
\quad 1\ i\ a_2\cdots a_{k-1}\\
\quad i\ a_1\ a_2\cdots a_{k-1}\\
(i-1)a_1\ a_2\cdots a_{k-1}\\
\qquad\qquad\vdots\\
\quad 4\ a_1\ a_2\cdots a_{k-1}.
\end{array}$$

Case 3. $n = 2$. Let the v treatments be $1, \ldots, i, a_1, \ldots, a_{k-1}, b_1, \ldots, b_{k-1}$. The blocks in $S_0 \cup S_1$ of Case 2 plus the following sequence S_2 of blocks shows that $f'(v, k) \leqq d(v, k)$ with $a_{k-3}b_{k-1}$ as the hole:

$$\begin{array}{l}
b_1\ a_1\ a_2\cdots a_{k-1}\\
b_2\ a_1\ a_2\cdots a_{k-1}\\
b_2\ b_3\ a_2\cdots a_{k-1}\\
\qquad\qquad\vdots\\
\qquad b_2\cdots b_{k-1}\ a_{k-2}\ a_{k-1}\\
\qquad b_2\cdots b_{k-1}\ 3\ a_{k-1}\\
3\ b_1\ b_2\cdots b_{k-1}\\
2\ b_1\ b_2\cdots b_{k-1}\\
1\ b_1\ b_2\cdots b_{k-1}\\
4\ b_1\ b_2\cdots b_{k-1}\\
\qquad\qquad\vdots\\
i\ b_1\ b_2\cdots b_{k-1}\\
a_1\ b_1\ b_2\cdots b_{k-1}\\
\qquad\qquad\vdots\\
a_{k-4}\ b_1\ b_2\cdots b_{k-1}.
\end{array}$$

Case 4. $n \geqq 3$. Let the v treatments be $1, \ldots, v_0, a_1, \ldots, a_{k-1}, b_1, \ldots, b_{k-1}, c_1, \ldots, c_{k-1}$, where $v = 3(k-1) + v_0$ and $v_0 \geqq k+1$. By induction, there exists an SCDH $(v_0, k, d(v_0, k)) =$ SCDH (V_0, S), with $1v_0$ as the hole, where $V_0 = \{1, 2, \ldots, v_0\}$. Without loss of generality, we assume the last block of S consists of $1, 2, \ldots, k$.

Let S'_1 and S'_2 be the sequences of blocks obtained from S_1 and S_2, respectively, by changing the symbol i to v_0. We now obtain a sequence of $d(v, k) = d(v_0, k) + 3v_0 + 6k - 12$ required blocks by using S, S'_1, S'_2 in that order, and appending the following blocks with $b_{k-3}c_{k-1}$ missing:

$$\begin{array}{l}
c_1\, b_1\, b_2 \cdots b_{k-1} \\
c_2\, b_1\, b_2 \cdots b_{k-1} \\
c_2\, c_3\, b_2 \cdots b_{k-1} \\
\qquad\vdots \\
c_2 \cdots c_{k-1}\, b_{k-2}\, b_{k-1} \\
c_2 \cdots c_{k-1}\, a_{k-3}\, b_{k-1} \\
c_1 \cdots c_{k-1}\, a_{k-3} \\
c_1 \cdots c_{k-1}\, a_{k-2} \\
c_1 \cdots c_{k-1}\, a_{k-1} \\
c_1 \cdots c_{k-1}\, a_{k-4} \\
\qquad\vdots \\
c_1 \cdots c_{k-1}\, a_1 \\
c_1 \cdots c_{k-1}\, b_1 \\
\qquad\vdots \\
c_1 \cdots c_{k-1}\, b_{k-4} \\
c_1 \cdots c_{k-1}\, 1 \\
\qquad\vdots \\
c_1 \cdots c_{k-1} v_0.
\end{array}$$

This shows that $f'(v, k) \leqq d(v, k)$ for any $v \geqq k + 1 \geqq 5$ by induction. □

By Theorem 2.1 and some elementary calculations, we can obtain

$$f'(v, k) \leqq \frac{v^2 + (k-5)v}{2(k-1)} - \frac{(k-3)(7k-1)-4}{8(k-1)},$$

which, together with Lemmas 1.1 and 1.3, immediately implies the following corollary.

COROLLARY 2.2. (i) $f'(v, 4) = \lceil (v^2 - v - 8)/6 \rceil$ *for all* $v \geqq 5$;

(ii) $\lceil (v^2 - v - 6)/6 \rceil \leqq f(v, 4) \leqq \lceil (v^2 - v - 2)/6 \rceil$ *for all* $v \geqq 5$. *In particular*, $f(v, 4) = (v^2 - v - 2)/6$ *if* $v \equiv 2 \pmod 3$.

3. A lower bound. Select a perfect SCD (V, B) such that $V = \{x_1, \ldots, x_v\}$ and $B = \{B_1, \ldots, B_b\}$. We say that x is *introduced in* B_i if $x \in B_i \backslash B_{i-1}$. For this purpose, we assume that $B_0 = \varnothing$, so that all members of B_1 are introduced in B_1. We say that x is *deleted from* B_i if $x \in B_{i-1} \backslash B_i$. Let s denote the smallest subscript such that $B_1 \cup \cdots \cup B_s = V$, and let T_i denote the set of treatments in V that are introduced exactly i times among the blocks $B_1, \ldots, B_s$, and $t_i = |T_i|$ for $0 \leqq i \leqq s$. Trivially, $t_0 = 0$ by the choice of s. Moreover, we have $v = \sum_{j=1}^{s} t_j$ and $s + k - 1 = \sum_{j=1}^{s} jt_j = v + \sum_{j=2}^{s} (j-1)t_j$. Therefore,

$$(1) \qquad s = v - k + 1 + \sum_{j=2}^{s} (j-1)t_j.$$

Now define

$$X = \{B_i \mid i > s \text{ and } B_i \backslash (B_s \cup \cdots \cup B_{i-1}) \neq \varnothing\},$$

$$Y = \{B_i \mid i > s \text{ and } B_{i-1} \backslash (B_i \cup \cdots \cup B_b) \neq \varnothing\}, \quad \text{and}$$

$$C = \{B_i \mid i > s \text{ and } B_i \notin X \cup Y\}.$$

Then we have the following result.

LEMMA 3.1. *It holds that* $|X| = |Y| = v - k$.

Proof. Assume that $\alpha \in B_s \backslash B_{s-1}$; then we have $\alpha \notin B_1 \cup \cdots \cup B_{s-1}$ by the choice of s. For any $x \notin B_s$, since the pair $x\alpha$ must appear together in some block, we see that $x \in B_{s+1} \cup \cdots \cup B_b$. This indicates that each treatment not in B_s is introduced in some block after B_s or, equivalently, $|X| = v - k$. Similarly, for any $x \notin B_b$, x is deleted from some block after B_s, which implies that $|Y| = v - k$. □

Write $T'_1 = \bigcup_{i=2}^{s} T_i$ and $t'_1 = |T'_1|$ for convenience. We have the following lemma.

LEMMA 3.2. *It holds that* $|X \cap Y| \leqq t'_1 + k$.

Proof. Assume that $X \cap Y = \{B_{r_1}, \ldots, B_{r_m}\}$, where $r_1 < r_2 < \cdots < r_m$ and $m \geqq t_1 + k + 1$. Also, assume that a_i and b_i are deleted from and introduced in B_{r_i}, respectively, for $1 \leqq i \leqq m$. Clearly, we have $a_i \neq a_j$ and $b_i \neq b_j$ whenever $i \neq j$ by the definitions of X and Y. We also note that, for any $1 \leqq i \leqq j \leqq m$, $a_i \neq b_j$, and the pair $a_i b_j$ must appear together in some block B_l, where $l < s$. Let $P_i = (a_i, b_i)$ for $1 \leqq i \leqq m$. We now obtain a new sequence $\{Q_i\}$ of ordered pairs based on $\{P_i\}$ by applying the following procedure.

Step 1. Select $1 \leqq j < l \leqq m$ such that $b_j = a_l \in T'_1$ if possible, delete P_l, and replace P_j by (a_j, b_l). Then denote the resulting sequence of ordered pairs by $\{P'_i\}$, where $1 \leqq i \leqq m - 1$. Similarly, select $1 \leqq j < l \leqq m - 1$, say $P'_j = (a'_j, b'_j)$ and $P'_l = (a'_l, b'_l)$, such that $b'_j = a'_l \in T'_1$ if possible, delete P'_l, and replace P'_j by (a'_j, b'_l). Then denote the resulting sequence of ordered pairs by $\{P''_i\}$, where $1 \leqq i \leqq m - 2$. We continue this process until we obtain a sequence $\{Q'_i\}$ of ordered pairs so that $Q'_i \cap Q'_j \cap T'_1 = \varnothing$ whenever $i \neq j$, where we consider each ordered pair Q'_i as a set consisting of two corresponding elements.

Step 2. For any i such that $Q'_i \cap T'_1 \neq \varnothing$, we delete Q'_i. Then denote the resulting subsequence of $\{Q'_i\}$ by $\{Q_i\}$, where $1 \leqq i \leqq r$, say.

Without loss of generality, we assume that $Q_i = (x_i, y_i)$ for $1 \leqq i \leqq r$. Then it can be easily checked that the sequence $\{Q_i\}$ has the following properties:

(i) $\{x_i, y_i\} \subseteq T_1$ for $1 \leqq i \leqq r$;

(ii) $x_i \neq x_j$ and $y_i \neq y_j$ whenever $i \neq j$;

(iii) For any $1 \leqq i \leqq j \leqq r$, we have $x_i \neq y_j$, and the pair $x_i y_j$ must appear together in some block B_l, where $l < s$;

(iv) $r \geqq m - t'_1 \geqq k + 1$.

Suppose that x_1 is introduced in B_{c_1}, where $c_1 \leqq s$. Let d_1 be the smallest index such that $x_1 \in B_{d_1}$ and $\{y_1, \ldots, y_r\} \subseteq B_{c_1} \cup B_{c_1+1} \cup \cdots \cup B_{d_1}$. Then we have $d_1 \leqq s$ and $x_1 \in B_{c_1} \cap B_{c_1+1} \cap \cdots \cap B_{d_1}$ because $x_1 \in T_1$. Since $r \geqq k + 1$, we can select $y_{i_1} \notin B_{c_1}$ for some $2 \leqq i_1 \leqq r$. Then, however, $x_2 y_{i_1}$ must appear together in some block B_l, where $c_1 < l \leqq s$, which implies that $x_2 \in B_{c_1+1} \cup \cdots \cup B_s$. Similarly, we can select $y_{j_1} \notin B_{d_1}$ for some $2 \leqq j_1 \leqq r$. This implies that $x_2 \in B_1 \cup \cdots \cup B_{d_1-1}$, since $x_2 y_{j_1}$ must appear together in some block B_l, where $1 \leqq l < d_1$. Therefore, we have $x_2 \in B_{c_1+1} \cup \cdots \cup B_{d_1-1}$ because $x_2 \in T_1$.

Now, assume that c_2 is the smallest index such that $c_2 \geqq c_1$ and $\{x_1, x_2\} \subseteq B_{c_2}$ and that d_2 is the largest index such that $d_2 \leqq d_1$ and $\{x_1, x_2\} \subseteq B_{d_2}$. This indicates that

$c_1 \leqq c_2 \leqq d_2 \leqq d_1$ and $\{x_1, x_2\} \subseteq B_{c_2} \cap \cdots \cap B_{d_2}$. Then we can easily check that $\{y_2, \ldots, y_r\} \subseteq B_{c_2} \cup \cdots \cup B_{d_2}$ by property (i) of the sequence $\{Q_i\}$.

Continuing this process, we can obtain a sequence of indices $c_1 \leqq c_2 \leqq \cdots c_k \leqq d_k \leqq \cdots \leqq d_1$, so that $\{x_1, \ldots, x_l\} \subseteq B_{c_l} \cap \cdots \cap B_{d_l}$ and $\{y_l, \ldots, y_r\} \subseteq B_{c_l} \cup \cdots \cup B_{d_l}$ for every $1 \leqq l \leqq k$. This, however, is impossible, since $r \geqq k + 1$ and each block has length k. This contradiction shows that $\lceil X \cap Y| \subseteq t'_1 + k$. □

We are now ready to prove the main result of this section.

THEOREM 3.3. *$f(v, k) \geqq 3v - 4k + 1$, with equality if $k \geqq 4$ and $2k + 1 \leqq v \leqq 3k - 1$.*

Proof. Applying (1) and Lemmas 3.1 and 3.2, we have

$$\begin{aligned} f(v, k) &= s + (b - s) \\ &= v - k + 1 + \sum_{j=2}^{s} (j - 1)t_j + |X \cup Y| + |C| \\ &\geqq v - k + 1 + \sum_{j=2}^{s} (j - 1)t_j + |X| + |Y| - |X \cap Y| \\ &= 3(v - k) + 1 - (|X \cap Y| - t'_1) + \sum_{i=3}^{s} (i - 2)t_i \\ &\geqq 3(v - k) + 1 - k \\ &= 3v - 4k + 1. \end{aligned}$$

In particular, if $k \geqq 4$ and $2k + 1 \leqq v \leqq 3k - 1$, we have $f(v, k) = 3v - 4k + 1$ by Theorem 2.1. This completes the proof of Theorem 3.3. □

By Lemma 1.2 and Theorem 3.3, we have that the smallest unknown number for $f(v, 4)$ is $v = 12$, in which case we proved that $21 \leqq f(12, 4) \leqq 22$ by Corollary 2.2.

Lemmas 1.1 and 1.2 and Theorem 3.3 immediately imply the following corollary.

COROLLARY 3.4. *$f(v, k) \geqq 2(v - k) + 1$, with equality if and only if $v \leqq 2k$ and $(v, k) \neq (4, 2)$.*

4. A recursive method.

THEOREM 4.1. *$f(mv, mk) \leqq [f(v, k) - 1]m + 1$ for any positive integer m.*

Proof. Select an SCD (v, k, b) on a v-set $X = \{1, 2, \ldots, v\}$ having block-sequence $B = \{B_1, B_2, \ldots, B_b\}$, where $b = f(v, k)$. Assume that $x_i \in B_i \backslash B_{i+1}$ and $y_i \in B_{i+1} \backslash B_i$ for $1 \leqq i \leqq b - 1$. Let $V = \bigcup_{i=1}^{v} V_i$ be a union of v disjoint sets V_i of size m and define

$$V_{x_i} = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{im}\}, \qquad V_{y_i} = \{\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}\},$$

$$B'_{im} = \bigcup_{j \in B_{i+1}} V_j \quad \text{for } 0 \leqq i \leqq b - 1,$$

$$B'_j = B'_{(i-1)m} \cup \{\beta_{i1}, \ldots, \beta_{ij^*}\} \backslash \{\alpha_{i1}, \ldots, \alpha_{ij^*}\}$$

$$\text{for } 1 \leqq i \leqq b - 1 \text{ and } 0 < j^* = j - (i - 1)m < m,$$

$$B' = (B'_0, B'_1, \ldots, B'_{(b-1)m}).$$

Therefore we have obtained an SCD (V, B') consisting of $(b - 1)m + 1$ blocks, which implies that $f(mv, mk) \leqq [f(v, k) - 1]m + 1$. □

By Theorems 3.3 and 4.1, we immediately have the following corollary.

COROLLARY 4.2. *If $f(v, k) = 3v - 4k + 1$, then $f(mv, mk) = 3mv - 4mk + 1$ for every positive integer m.*

THEOREM 4.3. *It holds that* $f(v+2, k+1) \leqq f(v,k)+2$.

Proof. Select an $\mathrm{SCD}(V, B) = \mathrm{SCD}(v, k, b)$ with block-sequence $B = (B_1, B_2, \ldots, B_b)$, where $b = f(v,k)$. Clearly, it suffices to construct an SCD $(v+2, k+1, b+2)$.

Let s be the smallest index such that $B_1 \cup \cdots \cup B_s = V$, and $V' = V \cup \{\alpha, \beta\}$. Select an arbitrary element x in B_s. Now define a sequence of $b+2$ new blocks $B' = (B'_1, \ldots, B'_{b+2})$ as follows:

$$B'_i = \begin{cases} B_i \cup \{\alpha\} & \text{if } 1 \leqq i \leqq s, \\ B_s \cup \{\alpha, \beta\} \setminus \{x\} & \text{if } i = s+1, \\ B_{i-2} \cup \{\beta\} & \text{if } i \geqq s+2. \end{cases}$$

Then the SCD (V', B') satisfies the requirement. □

By induction on m, Theorem 4.3 can be generalized as follows.

COROLLARY 4.4. $f(v+2m, k+m) \leqq f(v,k) + 2m$ *for all* $m \geqq 1$.

Similar to Corollary 4.2, we have the following corollary.

COROLLARY 4.5. *If* $f(v, k) = 3v - 4k + 1$, *then* $f(v+2m, k+m) = 3(v+2m) - 4(k+m) + 1$ *for every positive integer* m.

Trivially, we have $f(v, k+1) \leqq f(v,k) - 1$ and $f(v+1, k+1) \leqq f(v,k)$. In general, we have the following conjecture.

Conjecture 4.1. $f(v+n, k+1) \leqq f(v,k) + \binom{n+1}{2} - 1$ for all $n \geq 0$.

By induction on m, we have that Conjecture 4.1 is equivalent to the following conjecture.

Conjecture 4.2. $f(v+mn, k+m) \leqq f(v,k) + m[\binom{n+1}{2} - 1]$ for all nonnegative integers m and n.

REFERENCES

[1] J. D. HORTON, *Sub-latin squares and incomplete orthogonal arrays*, J. Combin. Theory, 16 (1974), pp. 23–33.

[2] W. H. MILLS AND R. C. MULLIN, *Packings and coverings*, in Contemporary Design Theory—A Collection of Surveys, John Wiley, New York, 1992.

[3] D. R. STINSON, *Frames for Kirkman triple systems*, Discrete Math., 65 (1987), pp. 289–300.

[4] W. D. WALLIS, J. L. YUCAS, AND G.-H. ZHANG, *Single-change covering designs*, Designs, Codes and Cryptography, 3 (1993), pp. 9–19.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 172–183, May 1994

ON THE ANGULAR RESOLUTION OF PLANAR GRAPHS*

SETH MALITZ† AND ACHILLEAS PAPAKOSTAS‡

Abstract. It is a well-known fact that every planar graph admits a planar straight-line drawing. The *angular resolution* of such a drawing is the minimum angle subtended by any pair of incident edges. The *angular resolution* of the graph is the supremum angular resolution over all planar straight-line drawings of the graph. In a recent paper by Formann et al. [*Proc.* 31*st IEEE Sympos. on Found. of Comput. Sci.*, 1990, pp. 86–95], the following question is posed: Does there exist a constant $r(d) > 0$ such that every planar graph of maximum degree d has angular resolution $\geq r(d)$ radians? The present authors show that the answer is yes and that it follows easily from results in the literature on disk-packings. The conclusion is that every planar graph of maximum degree d has angular resolution at least α^d radians, $0 < \alpha < 1$ a constant. In an effort to assess whether this lower bound is existentially tight (up to constant α), a very natural linear program (LP) that bounds the angular resolution of a planar graph the authors analyze from above. The optimal value of this LP is shown to be $\Omega(1/d)$, which suggests that the α^d lower bound might be improved to $\Omega(1/d)$. Although this matter remains unsettled for general planar graphs, $\Omega(1/d)$ is shown to be a lower bound on angular resolution for outerplanar graphs. Finally, an infinite family of triangulated planar graphs with maximum degree 6 is constructed such that exponential area is required to draw each member in planar straight-line fashion with angular resolution bounded away from zero.

Key words. planar graph, outerplanar graph, angular resolution, disk-packing, max-flow min-cut theorem

AMS subject classifications. 51, 68

1. Introduction.

Graphs are useful devices for representing many important objects in computer science, such as VLSI circuits, parallel computer architectures, networks of terminals, state graphs, data-flow graphs, Petri nets, entity-relationship diagrams, and so forth. When the structure of such a graph (signifying one of these objects) must be understood by a human being, a visual representation of the graph can be indispensable. Naturally, if there are many graphs under consideration, it is desirable that the process used to render them be automatic. The annotated bibliography of Eades and Tamassia [6] is a superb survey of graph-drawing algorithms and related theoretical results.

Perhaps the simplest way to render a graph is to represent the vertices as distinct points in the plane, and the edges as straight-line segments. There now remains the task of positioning the vertices so as to yield a drawing with low visual complexity. Naturally, it makes sense to optimize certain parameters defined over the space of all drawings. These may include minimizing the number of edge-crossings, maximizing uniformity in the placement of vertices, maximizing "angular spread" among incident edges, and so forth.

It is, in fact, this latter goal of attempting to generate large angular spread among incident edges that is the subject of the present paper. The first authors to formalize this concept and explore some of its properties were Formann et al. [8], who called this quantity the *resolution* of the graph. In this paper, we call it more precisely the *angular resolution.* Given an arbitrary graph G, the *angular resolution* of a straight-line drawing for G in the plane is defined to be the minimum angle subtended by any pair of edges incident to the same vertex. The *angular resolution* of G is then the supremum angular resolution over all straight-line drawings of G. (Thus, for example, if G has maximum

* Received by the editors September 11, 1992; accepted for publication (in revised form) January 11, 1993.

† Department of Computer Science, University of Massachusetts, Amherst, Massachusetts 01003 (smalitz@boulder.colorado.edu). This author's research was supported in part by National Science Foundation Grant CCR-8909953.

‡ Department of Computer Science, University of Texas at Dallas, Richardson, Texas 75083 (papakost@utdallas.edu).

degree d, then the angular resolution of G is bounded above by $2\pi/d$. In general, however, this bound is not achieved, as witnessed by the complete graph K_3, which has maximum degree 2 and angular resolution $\pi/3$.)

The authors of [8] provide many interesting results. They prove general upper and lower bounds on the angular resolution of a graph as a function of its maximum degree and show how to create straight-line drawings with high angular resolution for many interesting graph families. In particular, they show that any planar graph with maximum degree d has angular resolution $\Omega(1/d)$, independent of the number of vertices. However, in these drawings, line segments representing the edges of the planar graph are allowed to cross. This led the authors of [8] to pose the following question: Is there a constant $r(d) > 0$ such that any planar graph with maximum degree d has angular resolution $\geq r(d)$, where now angular resolution is computed over all *planar* straight-line drawings, i.e., drawings where the edges do not cross? (That every planar graph has a planar straight-line drawing is a famous theorem due independently to Steinitz and Rademacher [18], Wagner [23], Fàry [7], and Stein [17].)

This paper answers this question in the affirmative by showing that the angular resolution of any triangulated planar graph (i.e., a planar graph all of whose faces, including the exterior face, are triangles) with maximum degree d is at least α^d radians, where $0 < \alpha < 1$ is a constant. (From this, it follows that the angular resolution of an arbitrary planar graph of maximum degree d is bounded below by α^{3d}, since an arbitrary planar graph can be triangulated by at most tripling the degree at each vertex.) The proof is an easy application of two results from the literature on disk-packings. The first is that every triangulated planar graph can be realized as a disk-packing in the plane (for proofs, see Koebe [13], Andreev [1], [2], Thurston [20], and Colin de Verdiere [4], [5]). The second is the extension by Hansen [10] of the Rodin–Sullivan Ring Lemma (see Rodin and Sullivan [15]), which provides a tight lower bound on disk radii in a certain "flower-like" configuration of disks.

Next, we try to ascertain whether there exist triangulated planar graphs with maximum degree d and angular resolution exponentially small in d by analyzing a very natural family of linear programs (LPs) that are in one-to-one correspondence with the class of triangulated planar graphs. An LP in the family has a collection of variables called "angles." For a fixed triangulated planar graph G the associated LP requires us to maximize the minimum "angle" subject to the following constraints: (1) the "angles" at every vertex sum to 2π, (2) the "angles" inside any interior face sum to π, (3) the values of all "angles" are greater than or equal to zero. Clearly, any planar straight-line drawing of G satisfies these constraints. The converse is not necessarily so—not every solution to the constraints corresponds to a planar straight-line drawing. Consider, for instance, the complete graph K_4. Figure 1 shows an assignment to the "angles" that satisfies all the constraints, but that does not correspond to a planar straight-line drawing. So the above LP only partly captures the problem of computing angular resolution. More precisely, the value of this LP is an *upper bound* on the angular resolution of the associated graph. The question we attempt to answer is: How small can this upper bound be in terms of d, the maximum degree of the graph? By casting the LP as a max-flow problem and by applying the max-flow min-cut theorem, we prove that the optimum value of the LP is always $\Omega(1/d)$. This suggests that $\Omega(1/d)$, and not α^d, is perhaps the true lower bound on angular resolution for triangulated planar graphs with maximum degree d. We are currently unable to answer this question.

However, we do settle the matter for triangulated outerplanar graphs (i.e., outerplanar graphs in which every interior face is a triangle) with maximum degree d. We demonstrate not only that $\Omega(1/d)$ is a lower bound on angular resolution for this class of graphs, but

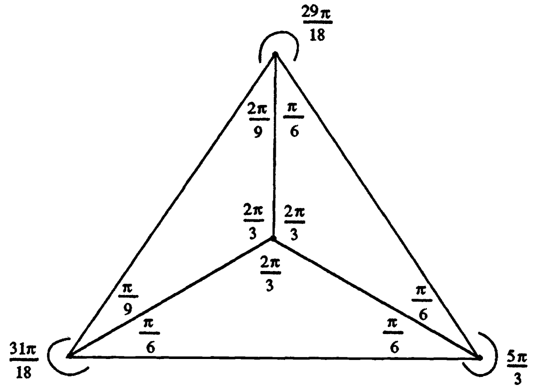


FIG. 1. *This assignment to the "angles" of K_4 satisfies the constraints of the linear program, but does not correspond to any planar straight-line drawing of K_4.*

that any such graph admits a planar straight-line drawing in which every angle is $\Omega(1/d)$ and all interior faces are *similar* isosceles triangles.

Finally, we consider the relationship between area and angular resolution in planar straight-line drawings of triangulated planar graphs. We construct a series triangulated planar graphs G_n of maximum degree 6 on $3n$ vertices such that, for any constant $c > 0$, any planar straight-line drawing of G_n (with vertices at gridpoints) with angular resolution greater than c requires area exponential in n. Thus, it is not always possible to achieve large angular resolution and small area in the same drawing. (Incidentally, de Fraysseix, Pach, and Pollack [9] and Schnyder [16] show that, if we ignore the issue of angular resolution, then any planar graph on n vertices admits a planar straight-line drawing in quadratic area, which is the best area possible in general.)

The organization of the paper is as follows. Section 2 describes recent related work of Kant [11], [12], who supplies, among other results, a linear-time algorithm that takes as input an arbitrary 3-connected, 3-valent planar graph and outputs a planar straight-line embedding of the graph in a hexagonal grid such that all but one interior face is convex. Section 3 demonstrates the α^d lower bound on angular resolution for triangulated planar graphs with maximum degree d. Section 4 analyzes a linear program that bounds the angular resolution of a triangulated planar graph from above and shows that its optimal value is always $\Omega(1/d)$. Section 5 establishes the lower bound of $\Omega(1/d)$ on angular resolution for triangulated outerplanar graphs. Finally, §6 discusses the issue of area versus angular resolution in planar straight-line drawings of triangulated planar graphs.

2. Related work. Recently, Kant [11], [12] has obtained a number of interesting results concerning planar straight-line drawings of planar graphs where angles are bounded away from zero and where the total number of bends in the edges is severely limited or zero. He has shown, for example, the following results:

(i) There is a simple linear-time algorithm that takes as input a 3-connected, 3-valent planar graph G on n vertices and outputs a planar embedding of G in a hexagonal grid (so that all angles are nonzero multiples $\pi/3$) of size $n/2 \times n/2$ such that at most one edge of G has a bend, and all but one interior face is convex. This drawing can be altered slightly so that there are no bends in any edge, all faces are convex, and all angles are $\geq \pi/6$ (although the new drawing is typically no longer an embedding in the hexagonal grid);

(ii) There is a simple linear-time algorithm that takes as input a connected 3-valent planar graph G and outputs a planar embedding of G in a hexagonal grid (so that all angles are nonzero multiples of $\pi/3$), where at most one edge of G has a bend. In some cases, the area required by such an embedding is exponential in the number of vertices in G. The drawing can be modified so that no edges are bent and all angles are $\geq \pi/6$.

Kant [11] conjectures that every 3-connected planar graph of maximum degree 4 can be embedded planarily in an octogonal grid (so that all angles are nonzero multiples of $\pi/4$) with at most a constant number of bends—at the moment, he has a linear time construction with $n/2$ bends in an $n \times n/2$ grid.

3. The angular resolution of planar graphs. A planar graph is *triangulated* if all of its faces, including the exterior face, are triangles. In this section, we show that, given any triangulated planar graph G with maximum degree d and distinguished face f_{ext}, there is a planar straight-line drawing of G with exterior face f_{ext} and angular resolution bounded below by α^d radians, where $0 < \alpha < 1$ is a constant. (It follows that the angular resolution of an arbitrary planar graph of maximum degree d is bounded below by α^{3d}, since an arbitrary planar graph can be triangulated by at most tripling the degree at each vertex.)

The proof follows easily from known results on disk-packings, which we describe below. First, however, we need some definitions. A *disk-packing* is a collection of disks in the plane with nonoverlapping interiors. A disk-packing P induces a planar graph G as follows: Place a vertex at the center of each disk and connect two vertices by an edge, if and only if their respective disks are tangent. If these edges are drawn as straight-line segments, then P induces a planar straight-line drawing of G. Let f_{ext} be the exterior face of G in this planar drawing. We say that P *realizes* the pair (G, f_{ext}).

THEOREM 3.1. *Given any triangulated planar graph G and distinguished face f_{ext}, there exists a disk-packing P that realizes (G, f_{ext}).*

A number of different proofs exist for this remarkable result (see Koebe [13], Andreev [1], [2], Thurston [20], and Colin de Verdiere [4], [5]). Very recently, Mohar [14] demonstrated a polynomial-time algorithm for obtaining such a disk-packing within any specified accuracy.

A finite sequence of closed disks having disjoint interiors is called a *chain* if each disk except the last is tangent to its successor. A chain is a *cycle* if the first and last disks are tangent. A *flower* with *n petals* consists of a cycle of n disks that surround the unit disk and are all tangent to it. Rodin and Sullivan [15] proved that there is a constant r depending only on n such that, for any flower with n petals, all petals have radius of at least r. They named this result the Ring Lemma. Hansen [10] went on to evaluate the largest value of r for which the Ring Lemma holds. The following is a trivial corollary of Hansen's result.

THEOREM 3.2. *There is a constant β between* 0 *and* 1 *such that, in any flower with n petals, every petal has radius of at least β^n.*

Given any triangulated planar graph G with maximum degree d and distinguished face f_{ext}, let P be a disk-packing that realizes (G, f_{ext}) such that the three outer disks of P all have unit radius. (That such a P exists can be seen by first applying Theorem 3.1 and then applying two bilinear transformations to obtain the three outer radii of unit value.) Let $\Gamma_P(G)$ denote the planar straight-line drawing of G induced by P. Let P' be the disk-packing obtained from P by adding nine additional unit disks to P such that each outer disk of P is now at the center of a flower with six unit disks as petals. Let G' be the planar graph induced by P'. Clearly, G' has maximum degree at most $d + 4$.

Consider any vertex v of G and consider all its neighbors in G'. The disks in P' that correspond to v and its neighbors form a flower (under suitable translation and adjustment for scale). By Theorem 3.2, all petals of this flower have radius of at least β^{d+4}, and thus it follows that all angles at v in $\Gamma_P(G)$ have value of at least α^d for some constant α between 0 and 1, independent of v.

Thus, we have the following theorem.

THEOREM 3.3. *Let G be an arbitrary triangulated planar graph with maximum degree d and distinguished face f_{ext}. Then G admits a planar straight-line drawing with exterior face f_{ext} and angular resolution of at least α^d radians, where α is a constant between* 0 *and* 1.

Hansen [10] constructs a flower F_d with d petals such that the smallest petal in the flower has radius bounded above by μ^d, for some constant μ between 0 and 1. The planar graph H_d induced by F_d is a triangulated planar graph of maximum degree d. It follows from Hansen's results that, for any disk-packing P that realizes H_d, the induced planar straight-line drawing $\Gamma_P(H_d)$ has angular resolution bounded above by ν^d, for some constant ν between 0 and 1. Hence, planar straight-line drawings induced by disk-packings cannot guarantee angular resolution of more than ν^d, in general.

4. A linear program and its evaluation.

In this section, we define a very natural LP associated with a triangulated planar graph G of maximum degree d whose optimal objective value is an upper bound on the angular resolution of G. By casting this LP as a max-flow problem and applying the max-flow min-cut theorem, we prove that its optimal value is at least $\pi/(3(d-1))$ for all G, as above.

Let $G = \langle V, E \rangle$ be a triangulated planar graph with maximum degree $d \geq 2$ and exterior face f_{ext}. Introduce a collection of variables called "angles," representing angle-values in a drawing of G. The LP that we are interested in maximizes the minimum "angle" subject to the following constraints: (1) the "angles" at every vertex sum to 2π, (2) the "angles" inside any interior face sum to π, and (3) all "angles" are greater than or equal to zero.

Special cases of this LP were considered for orthogonal drawings by Vijayan and Wigderson [22] and Tamassia [19], and for upward drawings by Bertolazzi and Di Battista [3]. Vijayan [21] considered the linear-programming problem of determining whether a planar graph with preassigned angles admits a planar straight-line realization. Our result is the following theorem.

THEOREM 4.1. *The above* LP *has optimal value of at least $\pi/(3(d-1))$.*

Proof. Let F be the collection of all faces of G, including f_{ext}. Let deg (v) denote the degree of vertex v in G. Let $V_{\text{ext}} \subseteq V$ be the subset of vertices that lie on f_{ext}. Let H denote the directed bipartite graph $\langle V, F, I \rangle$, where $(v, f) \in I$ if and only if vertex $v \in V$ and face $f \in F$ are incident in G.

We now describe a max-flow problem corresponding to the above LP. Begin with the digraph $H = \langle V, F, I \rangle$. Let all arcs of H have infinite capacity. Add to H two new vertices, s and t. Fix a real number $\alpha \in [0, \pi/d]$. Create an arc from s to each vertex $v \in V$ with capacity $2\pi - \alpha \deg(v)$. Create an arc from each vertex $f \in F - f_{\text{ext}}$ to vertex t with capacity $\pi - 3\alpha$. Create an arc from f_{ext} to t with capacity $5\pi - 3\alpha$. Call this capacitated network K. Clearly, K has an $s - t$ flow saturating all s-arcs, if and only if α is a lower bound for the LP.

Now, by the max-flow min-cut theorem, K has an $s - t$ flow of value

$$\sum_{v \in V} [2\pi - \alpha \deg(v)]$$

(and hence saturating all s-arcs) if and only if every $s - t$ cut has capacity $\geq$

$$\sum_{v \in V} [2\pi - \alpha \deg(v)].$$

Let us consider any finite-capacity $s - t$ cut. Such a cut is of the form $(S, \bar{S})$, where $S = \{s\} \cup V' \cup F'$, where $V' \subseteq V$, and $F' \subseteq F$ denotes a set of F-vertices containing all those incident to V'.

It suffices to consider only those cuts where V' and F' are *tightly coupled*, which means that they satisfy the following two properties: (1) F' consists of exactly those F-vertices incident to V', and (2) any superset of V' is incident to an F-vertex not in F'. Clearly, in the lowest-capacity cut $(S, \bar{S})$, the sets V' and F' are tightly coupled.

Thus, if S contains f_{ext}, we may assume S also contains at least one member of V_{ext}. However, if S contains f_{ext} and does not contain all members of V_{ext}, then a cut with lower capacity can be obtained by removing f_{ext} and all members of V_{ext} from S. Hence, in the cut $(S, \bar{S})$ with least capacity, S either contains no vertices among $V_{\text{ext}} \cup \{f_{\text{ext}}\}$ or contains all these vertices.

To begin, we assume the former, that S contains no vertices among $V_{\text{ext}} \cup \{f_{\text{ext}}\}$. In this case, the capacity of the cut $(S, \bar{S})$ is

$$\sum_{v \notin V'} [2\pi - \alpha \deg(v)] + \sum_{f \in F'} [\pi - 3\alpha].$$

We wish to determine the values of α for which this capacity is greater than or equal to $\sum_{v \in V} [2\pi - \alpha \deg(v)]$ (the saturation flow value for K) for all tightly coupled V', F', where $V' \subseteq V - V_{\text{ext}}$ and $F' \subseteq F - f_{\text{ext}}$. The inequality can be expressed as

$$\sum_{f \in F'} [\pi - 3\alpha] \geq \sum_{v \in V'} [2\pi - \alpha \deg(v)].$$

Simplifying further, we ask for the values of α where

$$(\pi - 3\alpha)|F'| - 2\pi|V'| + \alpha \deg(V') \tag{1}$$

is greater than or equal to zero for all tightly coupled V', F', where $V' \subseteq V - V_{\text{ext}}$ and $F' \subseteq F - f_{\text{ext}}$. (Here $\deg(V')$ denotes the sum of the degrees of all the vertices in V'.)

Henceforth, we denote the quantity (1) by $h(V', F')$.

We say a vertex (edge) is *incident* to a face of a planar graph if it participates in the cycle that defines the face. Let $G_{F'}$ be the subgraph of G whose vertices and edges are exactly those incident to faces in F'. Suppose that $G_{F'}$ consists of biconnected components $G_1, \ldots, G_p$. Let V_i' denote the vertices of G_i that belong to V' and let F_i' denote the faces of G_i that belong to F'. Then each pair V_i', F_i' is tightly coupled, and

$$h(V', F') = h(V_1', F_1') + \cdots + h(V_p', F_p').$$

Hence, a tightly coupled V', F' pair that makes $h(V', F')$ as small as possible is associated with a $G_{F'}$ that is biconnected. We now assume that $G_{F'}$ is biconnected.

Let V_b and E_b consist of all vertices and edges, respectively, in $G_{F'}$ that are incident to faces in $F - F'$. Clearly, V_b and V' are disjoint vertex sets. Call V_b the *boundary vertices* of $G_{F'}$ and E_b the *boundary edges* of $G_{F'}$. Note that each connected component of the graph (V_b, E_b) looks like a tree of cycles (by *cycle*, we always mean *simple* cycle). See Fig. 2(a).

Let T be such a tree of cycles. An articulation point v of T has *multiplicity m* if v participates in exactly m cycles of T. Consider the following transformation of $G_{F'}$. Duplicate each articulation point v of T a number of times equal to its multiplicity (maintaining each copy of v in V_b), so that T now becomes a cycle. See Fig. 2(b). This transformation in no way affects the values of $|V'|$, $|F'|$, and $\deg(V')$ and therefore preserves $h(V', F')$. Thus, without loss of generality, we assume that in $G_{F'}$ each connected

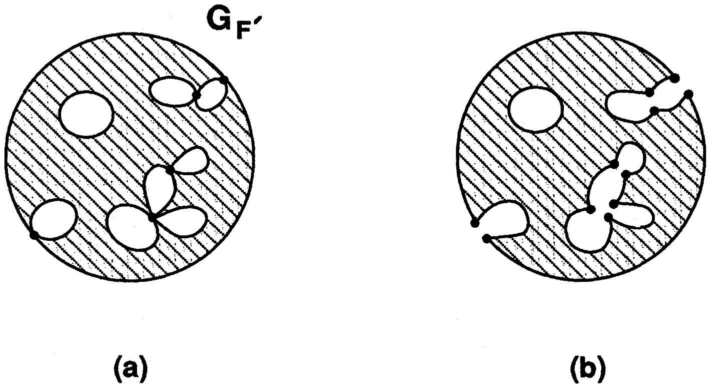


FIG. 2. (a) *Bold lines illustrate the connected components of* (V_b, E_b). *The faces of* $G_{F'}$ *that belong to* F' *are shaded.* (b) *What* $G_{F'}$ *looks like after duplicating all the articulation points of* (V_b, E_b).

component of (V_b, E_b) is a cycle. Let k be the number of such cycles that constitute (V_b, E_b).

We now derive a lower bound for $h(V', F')$ that holds for all tightly coupled V', F', where $V' \subseteq V - V_{\text{ext}}$ and $F' \subseteq F - f_{\text{ext}}$.

First, we observe that any triangulated planar graph on n vertices has $2n - 4$ faces. Thus

$$|F'| = 2(|V'| + |V_b| + k) - 4 - |V_b| = 2|V'| + |V_b| + 2k - 4,$$

where $|V_b|$ is the number of boundary vertices in $G_{F'}$. This formula is obtained by placing an imaginary vertex inside each cycle of (V_b, E_b) and triangulating.

In a triangulated planar graph on n vertices, the total vertex degree is $6n - 12$. From this, it follows that the total vertex degree in $G_{F'}$ is

$$\deg(V' \cup V_b) = 6(|V'| + |V_b| + k) - 12 - 2|V_b| = 6|V'| + 4|V_b| + 6k - 12.$$

Again, this formula is obtained by placing an imaginary vertex inside each cycle of (V_b, E_b) and triangulating.

Since every vertex in $G_{F'}$ has degree at most d, we have

$$\begin{aligned}\deg(V') &\geq \deg(V' \cup V_b) - d|V_b| = 6|V'| + 4|V_b| + 6k - 12 - d|V_b| \\ &= 6|V'| + (4 - d)|V_b| + 6k - 12.\end{aligned}$$

Thus

$$\begin{aligned}h(V', F') &= (\pi - 3\alpha)|F'| - 2\pi|V'| + \alpha \deg(V') \\ &\geq (\pi - 3\alpha)(2|V'| + |V_b| + 2k - 4) \\ &\quad - 2\pi|V'| + \alpha(6|V'| + (4 - d)|V_b| + 6k - 12) \\ &= (\pi - (d - 1)\alpha)|V_b| + 2\pi(k - 2).\end{aligned}$$

This last expression is smallest when $|V_b|$ and k assume their smallest possible values, namely when $V_b = 3$ and $k = 1$. Plugging in these values yields

$$h(V', F') \geq (\pi - (d - 1)\alpha)|V_b| + 2\pi(k - 2) \geq (\pi - (d - 1)\alpha)3 - 2\pi.$$

This last quantity is greater than or equal to zero if and only if $\alpha \leq \pi/3(d - 1)$. Thus, the condition $\alpha \leq \pi/3(d - 1)$ implies that $h(V', F') \geq 0$ for all tightly coupled V', F', where $V' \subseteq V - V_{\text{ext}}$ and $F' \subseteq F - f_{\text{ext}}$. This completes the case where S contains no members of $V_{\text{ext}} \cup \{f_{\text{ext}}\}$.

Let us now assume that S contains all the members of $V_{\text{ext}} \cup \{f_{\text{ext}}\}$. In this case, $f_{\text{ext}} \in F'$, and the capacity of the cut $(S, \bar{S})$ is

$$4\pi + \sum_{v \notin V'} [2\pi - \alpha \deg(v)] + \sum_{f \in F'} [\pi - 3\alpha].$$

We are interested in the values of α for which this capacity is always greater than or equal to $\sum_{v \in V} [2\pi - \alpha \deg(v)]$. That is, we ask for the values of α where

$$4\pi + (\pi - 3\alpha)|F'| - 2\pi|V'| + \alpha \deg(V')$$

is greater than or equal to zero for all $V' \subseteq V$ containing V_{ext} and $F' \subseteq F$ containing f_{ext}, where V' and F' are tightly coupled. Arguing as before, we obtain

$$h(V', F') \geq 4\pi + (\pi - (d-1)\alpha)|V_b| + 2\pi(k-2).$$

The right-hand side is as small as possible when $|V_b|$ and k are as small as possible, namely, when $|V_b| = 0$ and $k = 0$. Thus, we conclude that $h(V', F')$ is always greater than or equal to zero independent of α. This completes the case where S contains all the members of $V_{\text{ext}} \cup \{f_{\text{ext}}\}$.

We have now shown that all $s - t$ cuts $(S, \bar{S})$ have capacity $\geq \sum_{v \in V} [2\pi - \alpha \deg(v)]$ as long as $\alpha = \pi/(3(d-1))$. For such α, K has an $s - t$ flow that saturates all s-arcs. Thus, $\alpha = \pi/(3(d-1))$ is a lower bound for the optimal value of the LP. □

5. The angular resolution of outerplanar graphs. An outerplanar graph is *triangulated* if every interior face is a triangle. In this section, we prove that every triangulated outerplanar graph of maximum degree d has a planar straight-line drawing in which all angles are $\Omega(1/d)$ and where all interior faces are similar isosceles triangles.

First, we introduce some notation. Let AB denote a line segment in the plane with endpoints labeled A, B. Let ΔABC denote a triangle with vertices labeled A, B, C. Let $\angle ABC$ denote the *magnitude* of the angle between the line segments BA and BC. We begin with a technical lemma.

LEMMA 5.1. *Fix any integer $t \geq 2$. Let $\alpha = \pi/2(t+2)$. Draw ΔABC as an isosceles triangle with edge AB of length* 1 *lying on the x-axis, vertex C above the x-axis, and $\angle CAB = \angle CBA = \alpha$. Let s be the height of vertex C above the x-axis and let r be the length of the edge AC. Let R_A be a ray emanating from A and passing through a point Q_A directly above C at a height $l = s \sum_{i=0}^{\infty} r^i$ above the x-axis. Then the angle β between R_A and the edge AC satisfies* $(t-1)\alpha + \beta < \pi/2$. (See Fig. 3.)

Proof. The result follows from a short sequence of equalities and inequalities,

$$\begin{aligned}
\beta + \alpha &= \arctan 2l \\
&= \arctan \frac{2s}{1-r} < \frac{2s}{1-r} \qquad \text{(because } \tan x > x \text{ for } x > 0\text{)} \\
&= \frac{2 \sin \alpha}{2 \cos \alpha - 1} \qquad \text{(because } r \cos \alpha = \tfrac{1}{2} \text{ and } r \sin \alpha = s\text{)} \\
&< \frac{2\alpha}{2\left(1 - \frac{\alpha^2}{2}\right) - 1} = \frac{2\alpha}{1 - \alpha^2} \leq 4\alpha \qquad \left(\text{as long as } \alpha \leq \frac{1}{\sqrt{2}}\right).
\end{aligned}$$

Since $\alpha = \pi/2(t+2) < 1/\sqrt{2}$, we have from above, $\beta < 3\alpha$. Therefore, $(t-1)\alpha + \beta < (t+2)\alpha \leq \pi/2$. □

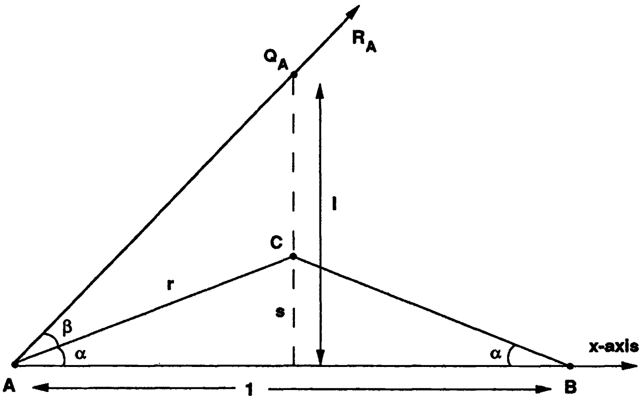


FIG. 3. *Construction for statement of Lemma* 5.1.

To state and prove the theorem of this section, we need two more definitions.

(i) Given an isosceles triangle ΔABC with axis of symmetry passing through C, define the *base angle* to be the value $\angle CAB$ (or, equivalently, $\angle CBA$).

(ii) Given a triangulated outerplanar graph G, the *dual graph* for G is obtained by placing a vertex in each interior face of G and connecting two vertices by an edge when the corresponding faces of G share an edge of G. Clearly, the dual graph of a triangulated outerplanar graph is always a tree.

THEOREM 5.1. *Let G be a triangulated outerplanar graph with exterior face F and maximum degree d. Let ΔABC be any face of G, where edge AB lies on F. Let δ_A denote the degree of vertex A in G and let δ_B denote the degree of vertex B in G. Then G admits a planar straight-line drawing $D(G)$ in which all interior faces are similar isosceles triangles with base angle $\alpha = \pi/2(d + 2)$. Assume that, in this drawing, the edge AB of triangle ΔABC is of length 1 and lies on the x-axis. Let β be the angle defined in Lemma 5.1. Then $D(G)$ lies inside a triangle ΔABP, where $\angle PAB \le (\delta_A - 1)\alpha + \beta < \pi/2$ and $\angle PBA \le (\delta_B - 1)\alpha + \beta < \pi/2$.*

Proof. Let T be the dual tree of G and suppose that T is rooted at ΔABC. We prove the theorem by induction on the height of T.

If the height of T is zero (so that G consists only of the face ΔABC), the theorem clearly holds.

Next, suppose that the theorem holds when the height of T is $\le h$. We now show that the theorem holds when the height of T equals $h + 1$.

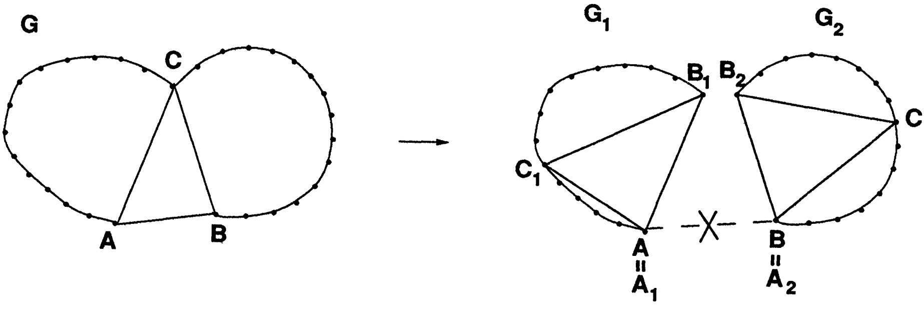


FIG. 4. *Construction for proof of Theorem* 5.1.

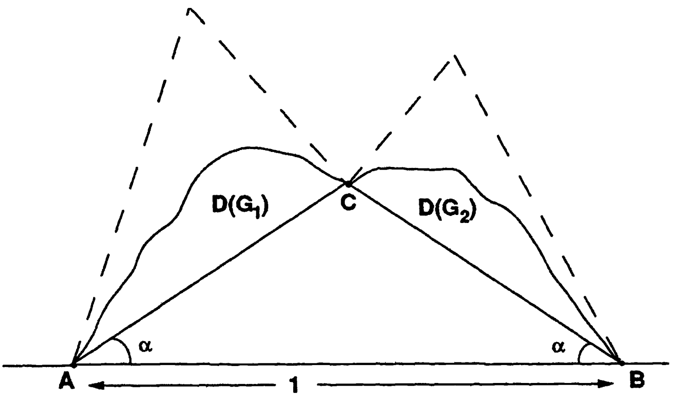


FIG. 5. *Construction for proof of Theorem* 5.1.

Consider the triangle ΔABC. By removing the edge AB from G and duplicating the vertex C (let B_1 and B_2 denote the copies of C), the graph G is split into two triangulated outerplanar subgraphs G_1 and G_2. See Fig. 4. Let $A_1 = A$ and $A_2 = B$. Let T_1 be the dual tree for G_1 rooted at the face $\Delta A_1B_1C_1$ and let T_2 be the dual tree for G_2 rooted at the face $\Delta A_2B_2C_2$. The height of both T_1 and T_2 is at most h. So, by the induction hypothesis, there are planar straight-line drawings $D(G_1)$ and $D(G_2)$ satisfying the conclusions of the theorem for each of G_1 and G_2, respectively.

Now we construct the drawing $D(G)$ for G and show that it satisfies the conclusions of the theorem. Take the naked triangle ΔABC and draw it isosceles with base angle α. Put the edge AB on the x-axis and give it length 1. Next, scale and rotate the drawings $D(G_1)$ and $D(G_2)$ so that they abut the triangle ΔABC, as shown in Fig. 5. Let $D(G)$ denote the resulting drawing of G. Clearly, $D(G)$ is planar because, by the induction hypothesis, each of $D(G_1)$ and $D(G_2)$ is planar and each lies inside a triangle of the appropriate shape so that there is no conflict between them.

It remains to show that $D(G)$ lies inside triangle ΔABP with the desired angles. Let $r < 1$ be the length of the edge AC in the triangle ΔABC. let s be the height of C above the x-axis. It is not difficult to see that the highest vertex in the drawing $D(G)$ has height less than $s \sum_{i=0}^{h-1} r^i < s \sum_{i=0}^{\infty} r^i$. Let l denote the value $s \sum_{i=0}^{\infty} r^i$. Consider the ray R_A emanating from A defined as follows. If $G_1 = \phi$, then R_A passes through a point Q_A directly above C at a height l above the x-axis. (The angle between R_A and AC is precisely β from Lemma 5.1.) If $G_1 \neq \phi$, then R_A coincides with the edge A_1P_1 of the triangle $\Delta A_1B_1P_1$ that contains $D(G_1)$. Similarly define the ray R_B. By the induction hypothesis, the angle between the ray R_A and the segment AB is at most $(\delta_A - 2)\alpha + \beta + \alpha = (\delta_A - 1)\alpha + \beta < \pi/2$. Similarly, the angle between the ray R_B and the segment BA is at most $(\delta_B - 2)\alpha + \beta + \alpha = (\delta_B - 1)\alpha + \beta < \pi/2$. Hence the two rays R_A and R_B cross at some point P, and $D(G)$ lies in the triangle ΔABP with the desired angles. □

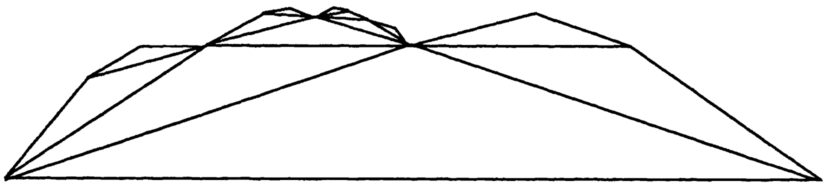


FIG. 6. *Planar straight-line drawing of a triangulated outerplanar graph using the method of Theorem* 5.1.

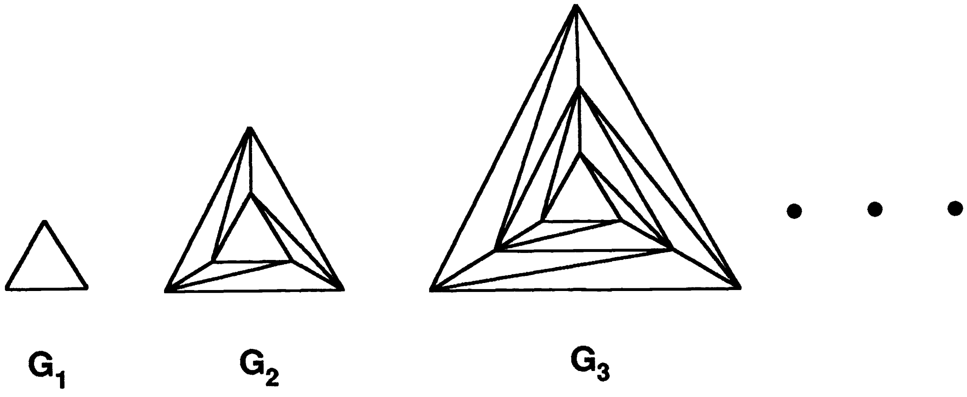


FIG. 7. *The graphs in the sequence G_n, $n = 1, 2, 3, \ldots$.*

Fig. 6 shows a planar straight-line drawing for a triangulated outerplanar graph using the method of Theorem 5.1.

6. Area versus angular resolution. Given a planar straight-line drawing of a planar graph in which the vertices of the graph reside at lattice points of a grid, the *area* of the drawing is simply the area of the smallest rectangle that contains it. The following theorem shows that, for certain bounded-degree triangulated planar graphs, it is not always possible to achieve small area and large angular resolution within the same planar straight-line drawing.

THEOREM 6.1. *Let G_n, $n = 1, 2, 3, \cdots$ be the sequence of triangulated planar graphs with maximum degree 6, illustrated in Fig. 7. Let c be any constant greater than zero. Then any planar straight-line drawing of G_n in which all angles have value greater than c requires area exponential in n.*

Proof. Let $A(n)$ be the area required by any planar straight-line drawing of G_n with angular resolution greater than c. It follows from the recursive structure of G_n and from elementary geometry that $A(n) = \Omega(A(n-1))$ (where the constant in the Ω depends on c). Thus, $A(n)$ is exponential in n. □

Acknowledgments. The authors thank Jerome Malitz, Eva Tardos, Peter Winkler, Oded Schramm, Gary Miller, Ron Rivest, Jin-Yi Cai, and Ken Stephenson for a variety of helpful discussions. Thanks also to the referee for a careful reading of an earlier draft and many helpful comments.

REFERENCES

[1] E. M. ANDREEV, *On convex polyhedra in Lobacevskii spaces*, Mat. Sb. (N. S.), 81 (1970), pp. 445–478; Math. USSR-Sb., 10 (1970), pp. 413–440. (English transl.)

[2] ———, *On convex polyhedra of finite volume in Lobacevskii space*, Mat. Sb. (N. S.), 83 (1970), pp. 256–260; Math. USSR-Sb., 12 (1970), pp. 255–259. (English transl.)

[3] P. BERTOLAZZI AND G. DI BATTISTA, *On upward drawing testing of triconnected digraphs*, in Proc. 7th ACM Sympos. on Computational Geometry, Conway, NH, 1991.

[4] Y. COLIN DE VERDIERE, *Empilements de cercles: convergence d'une methode de point fixe*, Forum Math., 1 (1989), pp. 395–402.

[5] ———, *Un principe variationnel pour les empilements de cercles*, Invent. Math., to appear.

[6] P. EADES AND R. TAMASSIA, *Algorithms for Drawing Graphs: An Annotated Bibliography*, Tech. Report CS-89-09, Dept. of Comput. Sci., Brown University, Providence, RI, 1989.

[7] I. FÀRY, *On straight-line representations of planar graphs*, Acta Sci. Math. (Szeged), 11 (1948), pp. 229–233.

[8] M. Formann, T. Hagerup, J. Haralambides, M. Kaufmann, F. T. Leighton, A. Simvonis, E. Welzl, and G. Woeginger, *Drawing graphs in the plane with high resolution*, in Proc. 31st IEEE Sympos. Found. of Comput. Sci., St. Louis, MO, 1990, pp. 86–95.

[9] H. de Fraysseix, J. Pach, and R. Pollack, *How to draw a planar graph on a grid*, Combinatorica, 10 (1990), pp. 41–51.

[10] L. Hansen, *On the Rodin and Sullivan ring lemma*, Complex Variables Theory Appl., 10 (1988), pp. 23–30.

[11] G. Kant, *Hexagonal Grid Drawings*, Tech. Report RUU-CS-92-06, Dept. Comput. Sci., Utrecht University, Utrecht, the Netherlands, 1992.

[12] ———, private communication, 1992.

[13] P. Koebe, *Kontaktprobleme auf der konformen Abbildung*, Ber. Verh. Saechs. Akad. Wiss. Leipzig, Math.-Phys. Kl., 88 (1936), pp. 141–164.

[14] B. Mohar, *Circle Packings of Maps in Polynomial Time*, Dept. Mathematics, Inst. of Mathematics, Physics and Mechanics, Univ. of Ljubljana, Jadranska 19, 61 111 Ljubljana, Slovenia, 1992, preprint.

[15] B. Rodin and D. Sullivan, *The convergence of circle packings to the Riemann mapping*, J. Differential Geom., 26 (1987), pp. 349–360.

[16] W. Schnyder, *Embedding planar graphs on the grid*, in Proc. 1st ACM-SIAM Sympos. on Discrete Algorithms, San Francisco, CA, 1990, p. 138–148.

[17] S. K. Stein, *Convex maps*, Proc. Amer. Math. Soc., 2 (1951), pp. 464–466.

[18] E. Steinitz and H. Rademacher, *Vorlesung Uber die Theorie der Polyeder* Springer, Berlin, 1934.

[19] R. Tamassia, *On embedding a graph in the grid with the minimum number of bends*, SIAM J. Comput., 16 (1987), pp. 421–444.

[20] W. P. Thurston, *The Geometry and Topology of 3-Manifolds*, Princeton Univ. Lecture Notes, Princeton, NJ.

[21] G. Vijayan, *Geometry of planar graphs with angles*, in Proc. 2nd ACM Sympos. on Computational Geometry, 1986, pp. 116–124.

[22] G. Vijayan and A. Wigderson, *Rectilinear graphs and their embeddings*, SIAM J. Comput., 14 (1985), pp. 335–372.

[23] K. Wagner, *Bemerkungen zum Viefarbenproblem*, Jber. Deutsch Math.-Verien, 46 (1936), pp. 26–32.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 184–193, May 1994

PLANAR SEPARATORS*

NOGA ALON†, PAUL SEYMOUR‡, AND ROBIN THOMAS§

Abstract. The authors give a short proof of a theorem of Lipton and Tarjan, that, for every planar graph with $n > 0$ vertices, there is a partition (A, B, C) of its vertex set such that $|A|$, $|B| < \frac{2}{3}n$, $|C| \leq 2(2n)^{1/2}$, and no vertex in A is adjacent to any vertex in B. Secondly, they apply the same technique more carefully to deduce that, in fact, such a partition (A, B, C) exists with $|A|$, $|B| < \frac{2}{3}n$, and $|C| \leq \frac{3}{2}(2n)^{1/2}$; this improves the best previously known result. An analogous result holds when the vertices or edges are weighted.

Key words. separators, k-shields, corner

1. The Lipton–Tarjan theorem. Our first objective is to give a short proof of the following theorem of Lipton and Tarjan [3] ($V(G)$ denotes the vertex set of the graph G):

(1.1). *Let G be a planar graph with $n > 0$ vertices. Then there is a partition (A, B, C) of $V(G)$ such that $|A|$, $|B| < \frac{2}{3}n$, $|C| \leq 2\sqrt{2}\sqrt{n}$, and no vertex in A is adjacent to any in B.*

Proof. We may assume that G has no loops or multiple edges, that $n \geq 3$, and (by adding new edges to G) that G is drawn in the plane in such a way that every region is bounded by a circuit of three edges. (Circuits have no "repeated" vertices.) Let $k = \lfloor\sqrt{2n}\rfloor$. For any circuit C of G, we denote by $A(C)$ and $B(C)$ the sets of vertices drawn inside C and outside C, respectively; thus $(A(C), B(C), V(C))$ is a partition of $V(G)$, and no vertex in $A(C)$ is adjacent to any in $B(C)$. Choose a circuit C of G such that

(i) $|V(C)| \leq 2k$,
(ii) $|B(C)| < \frac{2}{3}n$,
(iii) subject to (i) and (ii), $|A(C)| - |B(C)|$ is minimum.

(This is possible because the circuit bounding the infinite region satisfies (i) and (ii).)

We suppose, for a contradiction, that $|A(C)| \geq \frac{2}{3}n$. Let D be the subgraph of G drawn in the closed disc bounded by C. For $u, v \in V(C)$, let $c(u, v)$ (respectively, $d(u, v)$) be the number of edges in the shortest path of C (respectively, D) between u and v.

(1) $c(u, v) = d(u, v)$ *for all* $u, v \in V(C)$.

For certainly, $d(u, v) \leq c(u, v)$, since C is a subgraph of D. If possible, choose a pair $u, v \in V(C)$ with $d(u, v)$ minimum such that $d(u, v) < c(u, v)$. Let P be a path of D between u and v, with $d(u, v)$ edges. Suppose that some internal vertex w of P belongs to $V(C)$. Then

$$d(u, w) + d(w, v) = d(u, v) < c(u, v) \leq c(u, w) + c(w, v),$$

and so either $d(u, w) < c(u, w)$ or $d(w, v) < c(w, v)$; either case is contrary to the choice of u, v. Thus there is no such w. Let C, C_1, C_2 be the three circuits of $C \cup P$,

* Received by the editors April 30, 1991; accepted for publication (in revised form) February 10, 1993.

† Department of Mathematics, Sackler Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel. The research of this author was performed under a consulting agreement with Bellcore.

‡ Bellcore, 445 South Street, Morristown, New Jersey 07962.

§ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. This author was supported by Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers University, New Brunswick, New Jersey 08903.

where $|A(C_1)| \geq |A(C_2)|$. Now $|B(C_1)| < \frac{2}{3}n$, since

$$n - |B(C_1)| = |A(C_1)| + |V(C_1)|$$
$$> \tfrac{1}{2}(|A(C_1)| + |A(C_2)| + |V(P)| - 2) = \tfrac{1}{2}|A(C)| \geq \tfrac{1}{3}n.$$

However, $|V(C_1)| \leq |V(C)|$, since $|E(P)| \leq c(u, v)$, and so C_1 satisfies (i) and (ii). By (iii), $B(C_1) = B(C)$, and, in particular, $c(u, v) \leq 1$, which is impossible since $d(u, v) < c(u, v)$. This proves (1).

Suppose that $|V(C)| < 2k$. Choose $e \in E(C)$ and let P be the two-edge path of D such that the union of P and e forms a circuit bounding a region inside of C. Let v be the middle vertex of P and let P' be the path $C \backslash e$. Now $P \neq P'$, since $A(C) \neq \emptyset$, and so $v \notin V(C)$ by (1). Hence $P \cup P'$ is a circuit satisfying (i) and (ii), contrary to (iii). This proves that $|V(C)| = 2k$.

Let the vertices of C be $v_0, v_1, \ldots, v_{2k-1}, v_{2k} = v_0$, in order. There are $k + 1$ vertex-disjoint paths of D between $\{v_0, v_1, \ldots, v_k\}$ and $\{v_k, v_{k+1}, \ldots, v_{2k}\}$; for otherwise, by a well-known form of Menger's theorem for planar triangulations, there is a path of D between v_0 and v_k with $\leq k$ vertices, contrary to (1).

Let these paths be $P_0, P_1, \ldots, P_k$, where P_i has ends v_i, v_{2k-i} $(0 \leq i \leq k)$. By (1),

$$|V(P_i)| \geq \min(2i + 1, 2(k - i) + 1),$$

and so

$$n = |V(G)| \geq \sum_{0 \leq i \leq k} \min(2i + 1, 2(k - i) + 1) \geq \tfrac{1}{2}(k + 1)^2.$$

Yet $k + 1 > \sqrt{2n}$ by the definition of k, a contradiction. Thus our assumption that $|A(C)| \geq \frac{2}{3}n$ was false, and so $|A(C)| < \frac{2}{3}n$ and $(A(C), B(C), V(C))$ is a partition satisfying the theorem. □

2. Shields. In the remainder of the paper, we use the same technique more carefully to improve (1.1) numerically. A *separator* in a graph G is a partition (A, B, C) of $V(G)$ such that $|A|, |B| \leq \frac{2}{3}|V(G)|$ and no vertex in A is adjacent to any vertex in B; its *order* is $|C|$. Therefore, it is implied by (1.1) that any planar graph with n vertices has a separator of order $\leq 8^{1/2}n^{1/2}$, and we might try to find the smallest constant λ such that every planar graph with n vertices has a separator of order $\leq \lambda n^{1/2}$. The Lipton–Tarjan result (1.1) asserts that $\lambda \leq 8^{1/2} \simeq 2.828$, and this was improved by Gazit [2], who showed that $\lambda \leq \frac{7}{3} \simeq 2.333$. We give a further improvement, showing that $\lambda \leq \frac{3}{2} \cdot 2^{1/2} \simeq 2.121$. Incidentally, the best lower bound known appears to be that of Djidjev [1], who showed that

$$\lambda \geq \tfrac{1}{3}\sqrt{4\pi\sqrt{3}} \simeq 1.555.$$

Actually, we prove a slight strengthening, below (and indeed, we prove an extension when the vertices or edges have weights).

(2.1). *Let G be a loopless graph with n vertices, drawn in a sphere Σ. Then there is a simple closed curve F in Σ, meeting the drawing only in vertices, such that $n_1 + \frac{1}{2}n_3$, $n_2 + \frac{1}{2}n_3 \leq 2n/3$, and $n_3 \leq \frac{3}{2}(2n)^{1/2}$, where F passes through n_3 vertices and the two open discs bounded by F contain n_1 and n_2 vertices, respectively.*

We are concerned with graphs drawn in a disc or sphere Σ and, to simplify notation, we usually do not distinguish between a vertex of the graph and the point of Σ used in the drawing to represent the vertex, or between an edge and the open line segment representing it. A subset of Σ homeomorphic to the closed interval [0, 1] is called an *I-arc*. If G is drawn in Σ, a subset of Σ meeting the drawing only in vertices is *G-normal*.

The proof of (2.1) relies on the notion of a "k-shield." Let $k \geq 0$. A *k-shield* (in Δ) is a loopless graph G drawn in a closed disc Δ such that, denoting the boundary of Δ by $bd(\Delta)$,

(i) $|V(G) \cap bd(\Delta)| = k$,

(ii) $bd(\Delta)$ is G-normal, and

(iii) for every G-normal I-arc $F \subseteq \Delta$ with ends $x, y \in bd(\Delta)$, there is an I-arc $F' \subseteq bd(\Delta)$ with ends x, y such that $|V(G) \cap F'| \leq |V(G) \cap F|$.

We can view the proof of (1.1) as consisting of two parts (omitting the reduction to G being a planar triangulation, which is included only for convenience and can easily be avoided): Roughly, we show that, for any k, every planar graph either has a separator of order $\leq k$ or has a subgraph that is a k-shield; secondly, we show that any k-shield has at least approximately $\frac{1}{8}k^2$ vertices. Consequently, any planar graph with no separator of order $\leq k$ has at least approximately $\frac{1}{8}k^2$ vertices, and (1.1) follows.

We improve this as follows. First, $\frac{1}{8}$ is the wrong constant; we shall see that any k-shield has at least $\frac{1}{6}k^2$ vertices. ($\frac{1}{6}$ might not be the right constant, either.) Secondly, with a little care, we can find a k-shield in G that contains at most three-quarters of the vertices of G.

In this section, we prove the fact that any k-shield has at least $\frac{1}{6}k^2$ vertices and some related lemmas; these are applied to prove (2.1) in the next section.

The proof of the next result is due to A. Schrijver (in a private communication); our original proof was an application of a currently unpublished theorem of Randby about graphs drawn in the projective plane [5], but Schrijver's proof is simpler.

(2.2). *If G is a k-shield, then* $|E(G)| \geq \frac{1}{2}k(k-1)$.

Proof. We may assume that $k \geq 3$ and that G has no multiple edges. We may also assume that G is 2-connected, for, if not, then it includes a smaller k-shield, since one of the following holds:

(i) there is a closed disc $\Delta' \subseteq \Delta$ such that $bd(\Delta')$ is G-normal, $|bd(\Delta') \cap V(G)| \leq 1$, and the interior of Δ' meets G, but then we may produce a smaller k-shield by deleting the part of G in the interior of Δ';

(ii) there is a G-normal I-arc $F \subseteq \Delta$ with ends $x, y \in bd(\Delta) - V(G)$ and with $F \cap bd(\Delta) = \{x, y\}$ and $|F \cap V(G)| = 1$, dividing Δ into two closed discs Δ_1, Δ_2 such that the interiors of both Δ_1 and Δ_2 meet G, but then (since G is a k-shield, and we may assume that (i) is false) it follows that $|bd(\Delta_i) \cap V(G)| = 2$ for some i, say $i = 2$, and hence the restriction of G to Δ_1 is a smaller k-shield.

Thus, let G be 2-connected. It follows that there is a circuit C of G, bounding a closed disc in Δ that includes all the drawing of G. Let the vertices of G in $bd(\Delta)$ be $v_1, \ldots, v_k$, and, for $1 \leq i \leq k$, let l_i be the open line segment between v_i and v_{i+1} that is an arc-wise connected component of $bd(\Delta) - \{v_1, \ldots, v_k\}$ (where v_{k+1} means v_1). For $1 \leq i \leq k$, let r_i be the region of G in Δ including l_i. Then, for $1 \leq i \leq k$, the boundary of r_i consists of l_i together with a path from C, while every other region of G in Δ is an open disc and is bounded by a circuit of G. Let us say that a *corner* of G is a pair (v, r), where $v \in V(G)$ and r is a region of G in Δ incident with v. For any corner (v, r), there are precisely two edges of G incident with both v and r unless $r = r_i$ and $v = v_i$ or v_{i+1} for some i, when there is only one such edge. We call any such edge an *arm* of the corner.

We wish to define a new graph G' drawn in Δ. For each $e \in E(G)$, let x_e be a point of the open line segment representing e in the drawing of G. For $1 \leq i \leq k$, let a_i, b_i be distinct points of l_i, so that v_i, a_i, b_i, v_{i+1} occur in order. The vertex set of G' will be

$$\{a_1, b_1, a_2, b_2, \ldots, a_k, b_k\} \cup \{x_e : e \in E(G)\}.$$

The edges of G' correspond to the corners of G. For each corner (v, r) with two arms e, f, there is an edge of G' with ends x_e, x_f, drawn within r. For each corner (v, r) with one arm e, let $r = r_i$; then, if $v = v_i$, the corresponding edge of G' has ends a_i, x_e, while, if $v = v_{i+1}$, it has ends $b_i x_e$; in either case, it is drawn within r. This defines G' and its drawing. We see that every vertex of G' has valency 4, except for $a_1, b_1, \ldots, a_k, b_k$, which all have valency 1. Moreover, each region of G' in Δ either includes a (unique) vertex of G or is a subset of a region of G in Δ; every edge of G' is incident with one region of each type.

(1). *Let $F' \subseteq \Delta$ be an I-arc with ends $s, t \in bd(\Delta)$, not passing through any vertex of G'; and let F_1, F_2 be the two I-arcs in $bd(\Delta)$ with ends s, t. Then the number of edges of G' crossed by F' is at least* $\min(|F_1 \cap V(G')|, |F_2 \cap V(G')|)$.

For we may assume (by rerouting F') that $F' \cap r$ is an open line segment or null, for every region r of G' in Δ. As we traverse F' from s to t, the regions of G' we pass through correspond alternately to vertices and regions of G, and there is a G-normal I-arc F in Δ, passing through the same sequence of vertices and regions. Moreover, we may assume that F and F' have the same ends. Hence, F passes through at least $\min(|F_1 \cap V(G)|, |F_2 \cap V(G)|)$ vertices of G, since G is a k-shield; say $|F \cap V(G)| \geq |F_1 \cap V(G)|$. If both ends of F are in $V(G)$, then

$$|F' \cap E(G')| \geq 2|F \cap V(G)| - 2 \geq 2|F_1 \cap V(G)| - 2 = |F_1 \cap V(G')|,$$

and a similar computation applies if one or neither end of F is in $V(G)$. This proves (1).

Let us renumber $a_1, b_1, a_2, b_2, \ldots, a_k, b_k$ as $s_1, s_2, \ldots, s_k, t_1, t_2, \ldots, t_k$, respectively. From (1) and the result of [4], it follows that there are k mutually edge-disjoint paths $P_1, \ldots, P_k$ of G' joining s_i and t_i $(1 \leq i \leq k)$, respectively. Since for $1 \leq i < j \leq k$, P_i and P_j have a common vertex (because they must cross somewhere) and this vertex belongs to no other of the k paths, we deduce that G' has at least $\frac{1}{2}k(k-1)$ vertices of valency 4. Consequently, $|E(G)| \geq \frac{1}{2}k(k-1)$, as required. □

A k-shield G in Δ is *stable* if for every I-arc $L \subseteq bd(\Delta)$ with ends x, y and with $L \cap V(G) = \{x, y\}$, there is no edge e of G with ends x, y such that $L \cup e$ bounds a region of G in Δ.

(2.3). *If $k \geq 3$ and G is a stable k-shield, then* $|V(G)| \geq \frac{1}{6}k^2 + \frac{1}{2}k + 1$.

Proof. Let G be drawn in Δ. If some region of G in Δ is bounded by a two-edge circuit, we may delete one of these two edges. By continuing this process, we may assume there is no such region.

Let the vertices of G drawn in $bd(\Delta)$ be $v_1, \ldots, v_k$ in order. Add to G a new vertex v_0, edges with ends v_0, v_i $(1 \leq i \leq k)$ and edges with ends v_i, v_{i+1} $(1 \leq i \leq k)$, where v_{k+1} means v_1. We obtain a new planar graph G', with $|V(G')| = |V(G)| + 1$ and $|E(G')| = |E(G)| + 2k$. Moreover, G' can be drawn in a sphere so that no region has boundary consisting of a one- or two-edge circuit. Since $|V(G')| \geq 3$, it follows that $|E(G')| \leq 3|V(G')| - 6$, and hence

$$|E(G)| + 2k \leq 3(|V(G)| + 1) - 6.$$

By (2.2), however, $|E(G)| \geq \frac{1}{2}k(k-1)$, and the result follows. □

Similarly, for $k \geq 2$, any k-shield has $\geq \frac{1}{6}k^2 + \frac{1}{6}k + 1$ vertices. We do not know if the term $\frac{1}{6}k^2$ here is the best possible.

(2.4). *Let G be a graph drawn in a closed disc Δ, such that $|V(G) \cap bd(\Delta)| = k$ and $bd(\Delta)$ is G-normal. Suppose that, for every G-normal I-arc $F \subseteq \Delta$ with ends x, $y \in bd(\Delta)$ and $F \cap bd(\Delta) = \{x, y\}$, there is an I-arc $F' \subseteq bd(\Delta)$ with ends x, y such that $|F' \cap V(G)| \leq |F \cap V(G)|$. Then G is a k-shield.*

Proof. For distinct $x, y \in bd(\Delta)$, let

$$d(x, y) = \min(|F_1 \cap V(G)|, |F_2 \cap V(G)|),$$

where F_1, F_2 are the two I-arcs in $bd(\Delta)$ with ends x, y. We must show that $|F \cap V(G)| \geq d(x, y)$ for every G-normal I-arc $F \subseteq \Delta$ with ends $x, y \in bd(\Delta)$. We may assume that $F \cap bd(\Delta) \subseteq V(G) \cup \{x, y\}$ and we proceed by induction on $|F \cap bd(\Delta) - \{x, y\}|$. If this quantity is zero, the result follows from the hypothesis. Otherwise, there exists $z \in (F - \{x, y\}) \cap bd(\Delta)$, and $z \in V(G)$. Let $F_1, F_2 \subseteq F$ be the I-arcs with ends x, z and z, y, respectively. From the inductive hypothesis, $|F_1 \cap V(G)| \geq d(x, z)$ and $|F_2 \cap V(G)| \geq d(z, y)$. However,

$$|F \cap V(G)| = |F_1 \cap V(G)| + |F_2 \cap V(G)| - 1,$$

and $d(x, y) \leq d(x, z) + d(z, y) - 1$ (since $z \in V(G)$). The result follows. □

Let us say a *strong* k-shield is a graph G drawn in a closed disc Δ with $|V(G) \cap bd(\Delta)| = k$ and with $bd(\Delta)$ G-normal, such that, for every G-normal I-arc $F \subseteq \Delta$ with ends $x, y \in bd(\Delta)$ and with $F \cap bd(\Delta) = \{x, y\}$, either

(i) there is an I-arc $F' \subseteq bd(\Delta)$ with ends x, y such that $|F' \cap V(G)| < |F \cap V(G)|$, or

(ii) one of the two closed discs into which F divides Δ includes all of the drawing of G.

From (2.4), we see that every strong k-shield is a k-shield.

(2.5). *For $k \geq 3$, if G is a strong k-shield, then $|V(G)| \geq \frac{1}{6}k^2 + \frac{5}{6}k + \frac{2}{3}$.*

Proof. We may assume that G has no multiple edges. Since $k \geq 3$, it follows that no two vertices in $bd(\Delta)$ are adjacent (if an edge has both ends in $bd(\Delta)$, then we may choose F to violate conditions (i) and (ii) in the definition of strong k-shield, with the same ends as e and otherwise disjoint but "next to" e). Let the vertices of G in $bd(\Delta)$ be $v_1, \ldots, v_k$ in order. Let r be the region incident with v_1 and v_k (it is unique). Since G is a k-shield, r is not incident with any of $v_2, \ldots, v_{k-1}$. Let $u \neq v_1, v_k$ be incident with r (this exists since v_1, v_k are not adjacent). Add a new vertex v_{k+1} to G and an edge e with ends u, v_{k+1}, forming G'. Draw v_{k+1} in $r \cap bd(\Delta)$ and draw e within $r - bd(\Delta)$.

(1). *G' is a $(k + 1)$-shield.*

For let $F \subseteq \Delta$ be a G'-normal I-arc with ends $x, y \in bd(\Delta)$, and with $F \cap bd(\Delta) = \{x, y\}$. Let F_1, F_2 be the two I-arcs in $bd(\Delta)$ with ends x, y. We claim that

$$|F \cap V(G')| \geq \min(|F_1 \cap V(G')|, |F_2 \cap V(G')|).$$

If $|F \cap V(G)| > |F_1 \cap V(G)|$, then our claim holds, since

$$|F \cap V(G')| \geq |F \cap V(G)| \geq |F_1 \cap V(G)| + 1 \geq |F_1 \cap V(G')|.$$

We assume then that $|F \cap V(G)| \leq |F_i \cap V(G)|$ for $i = 1, 2$. Since G is a strong k-shield, we may assume that the closed disc in Δ bounded by $F \cup F_2$ includes the drawing of G. Hence $F_1 \cap V(G) \subseteq F \cap V(G)$. If $v_{k+1} \notin F_1$, then

$$|F_1 \cap V(G')| = |F_1 \cap V(G)| \leq |F \cap V(G)| \leq |F \cap V(G')|,$$

as required. If $v_{k+1} \in F_1$, then either v_{k+1} or u belongs to F (since u belongs to the closed disc bounded by $F \cup F_2$). In the first case, $v_{k+1} = x$ or y, and so

$$|F_1 \cap V(G')| = |F_1 \cap V(G)| + 1 \leq |F \cap V(G)| + 1 \leq |F \cap V(G')|.$$

In the second case, $u \in F \cap V(G)$ and $u \notin F_1 \cap V(G)$, and so

$$|F_1 \cap V(G')| = |F_1 \cap V(G)| + 1 \leq |F \cap V(G)| \leq |F \cap V(G')|.$$

This proves our claim that

$$|F \cap V(G')| \geq \min(|F_1 \cap V(G')|, |F_2 \cap V(G')|).$$

Consequently, G' is a $(k + 1)$-shield. This proves (1).

Certainly G' is a stable $(k + 1)$-shield, and so, from Theorem 2.3, we deduce that $|V(G')| \geq \frac{1}{6}(k + 1)^2 + \frac{1}{2}(k + 1) + 1$. Since $|V(G')| = |V(G)| + 1$, the result follows. □

3. The main argument. In § 1 we were concerned with the problem of finding a small cutset, defined by a simple closed curve, so that both sides of it contain about the same number of vertices. We can also give the vertices or edges weights and ask for a small cutset, defined by a simple closed curve, so that both sides contain about the same total weight. This is a little more complicated; for instance, although the analogue of (1.1) holds (that is, $2(2n)^{1/2}$), Gazit's proof of $\frac{7}{3}n^{1/2}$ does not extend, and until now $2(2n)^{1/2}$ was the best known. However, we show in (3.9) that, for any planar G and for any constant $\lambda \geq 2$, if $\lambda n^{1/2}$ works for the unweighted case, then it also works for the weighted case. In particular, our result of $\frac{3}{2}(2n)^{1/2}$ works for the weighted case.

A convenient common generalization of the different ways to assign weights to a planar graph is via "majorities." Let G be a graph drawn in a sphere Σ. A *noose* is a G-normal, simple closed curve $F \subseteq \Sigma$, and its *length* is $|F \cap V(G)|$. A *majority of order* k, where $k \geq 0$ is an integer, is a function "big" that assigns to every noose F of length $\leq k$ a closed disc big$(F) \subseteq \Sigma$ bounded by F satisfying the following two axioms:

Axiom 1. If $x, y \in \Sigma$ are distinct, and F_1, F_2, F_3 are G-normal I-arcs each between x and y and otherwise disjoint, and $F_1 \cup F_2$, $F_1 \cup F_3$, $F_2 \cup F_3$ all have length $\leq k$, and big$(F_1 \cup F_2)$ includes F_3, then big$(F_1 \cup F_2)$ includes one of big$(F_1 \cup F_3)$, big$(F_2 \cup F_3)$.

Axiom 2. If F is a noose with length $\leq \min(2, k)$, then either big$(F) - F$ contains a vertex of G or big(F) includes at least two edges of G.

This is connected with the weighted separator problem via the next two results. $\mathbf{R}_+$ denotes the set of nonnegative real numbers, and if $w : X \rightarrow \mathbf{R}_+$ is a function and $Y \subseteq X$, we denote $\Sigma(w(x) : x \in Y)$ by $w(Y)$.

(3.1). *Let G be a graph drawn in a sphere, let $w : V(G) \rightarrow \mathbf{R}_+$ be a function, and let $k \geq 0$ be an integer. Suppose that there is no noose F of length $\leq k$ such that*

$$w((D - F) \cap V(G)) + \tfrac{1}{2}w(F \cap V(G)) \leq \tfrac{2}{3}w(V(G))$$

for both closed discs D bounded by F. Then G has a majority of order k.

Proof. For each noose F of length $\leq k$, let big(F) be the (unique) closed disc D bounded by F such that

$$w((D - F) \cap V(G)) + \tfrac{1}{2}w(F \cap V(G)) > \tfrac{2}{3}w(V(G)).$$

The axioms may easily be verified. □

If G is drawn in Σ and $D \subseteq \Sigma$ is a closed disc with *bd*(D) G-normal, we denote the subgraph of G drawn in D by $G \cap D$.

(3.2). *Let G be a graph in a sphere, let $w : E(G) \rightarrow \mathbf{R}_+$ be a function, and let $k \geq 0$ be an integer. Suppose that*

(i) *$w(f) \leq \frac{2}{3}w(E(G))$ for each $f \in E(G)$, and*

(ii) *there is no noose F of length $\leq k$ such that $w(E(G \cap D)) \leq \frac{2}{3}w(E(G))$ for both closed discs D bounded by F.*

Then G has a majority of order k.

Proof. For each noose F of length $\leq k$, let big(F) be the unique closed disc D bounded by F with $w(E(G \cap D)) > \frac{2}{3}w(E(G))$. Again, the axioms may easily be verified. □

Let big be a majority of order k in G. A noose F is *optimal* if

(i) it has length $\leq k$;
(ii) subject to (i), $G \cap \text{big}(F)$ is minimal; and
(iii) subject to (i) and (ii), $|F \cap V(G)|$ is maximum.

(3.3). *Let G be a loopless graph drawn in a sphere Σ, let* big *be a majority of order $k \geq 0$, and let F be an optimal noose. Then $G \cap$ big(F) is a strong stable k-shield in* big(F).

Proof. Let $|F \cap V(G)| = k'$. We claim first that $G \cap \text{big}(F)$ is a strong k'-shield in big(F). Let big$(F) = \Delta$, let $F_3 \subseteq \Delta$ be a G-normal I-arc with ends $x, y \in F$ and with $F_3 \cap F = \{x, y\}$, and let F_1, F_2 be the two arcs between x, y in F. Suppose that

$$|F_3 \cap V(G)| \leq |F_1 \cap V(G)|, \ |F_2 \cap V(G)|.$$

Let $\Delta_i \subseteq \Delta$ be the closed disc bounded by $F_i \cup F_3$ $(i = 1, 2)$. We must show that one of Δ_1, Δ_2 includes $G \cap \Delta$. For $i = 1, 2$, $F_i \cup F_3$ is a G-normal noose with length $\leq k$, since

$$|(F_i \cup F_3) \cap V(G)| \leq |F \cap V(G)| \leq k.$$

From Axiom 1, we may assume that $\Delta_1 = \text{big}(F_1 \cup F_3)$. Since F is optimal, it follows that $G \cap \Delta_1 = G \cap \Delta$; that is, Δ_1 includes $G \cap \Delta$, as required. Thus, $G \cap \text{big}(F)$ is a strong k'-shield.

We claim that $G \cap \text{big}(F)$ is a stable k'-shield. This is clear if $k' \geq 3$ because every strong k'-shield with $k' \geq 3$ is stable, but needs proof if $k' \leq 2$. Suppose that e is an edge of $G \cap \Delta$ with ends $x, y \in bd(\Delta)$ and that $L \subseteq F$ is an I-arc with ends x, y such that $e \cup L$ bounds a region of G in Δ. Let $F_3 \subseteq \Delta$ be a G-normal I-arc with ends x, y, just on the other side of e from L, in the natural sense. From Axiom 2, big$(F_3 \cup L) \not\subseteq \Delta$, and so, from Axiom 1, big$(F_3 \cup (F - L)) \subseteq \Delta$, contrary to the optimality of F. Thus, $G \cap \Delta$ is a stable k'-shield in Δ.

Finally, we claim that $k' = k$. Suppose that $k' < k$ and let r be a region of $G \cap \Delta$ in Δ with $F \cap r \neq \emptyset$. Suppose that $v \in V(G \cap \Delta)$ is incident with r, and $v \notin F$. Choose distinct $x, y \in r \cap F$ and let F_3 be an I-arc with ends x, y and $F_3 \cap F = \{x, y\}$ and $F_3 \subseteq r \cup \{v\}$, passing through v. Since $k > k' \geq 0$, it follows that $|(F_1 \cup F_3) \cap V(G)| \leq 1 \leq k$, where $F_1 \subseteq r \cap F$ is an I-arc between x, y, and $|(F_2 \cup F_3) \cap V(G)| \leq k' + 1 \leq k$, where $F_2 \subseteq F$ is the other I-arc between x, y. By Axiom 2, big$(F_1 \cup F_3) \not\subseteq \Delta$, and so, by Axiom 1, big$(F_2 \cup F_3) \subseteq \Delta$, contrary to the optimality of F.

Hence, every $v \in V(G \cap \Delta)$ incident with r belongs to F. Since $G \cap \Delta$ is a k'-shield, it follows that $r \cap F$ is connected. If $F \subseteq r$, then r is incident with no vertex of $G \cap \Delta$, and so $G \cap \Delta$ is null, contrary to the second axiom. Hence, $r \cap F$ is an open line segment, with ends $x, y \in V(G)$. Since $G \cap \Delta$ is a k'-shield, it follows that r is incident with no vertex of $G \cap \Delta$ except x and y. In particular, $x \neq y$ since G is loopless, and there is an edge of G with ends x and y, incident with r. This is impossible, however, since $G \cap \Delta$ is a stable k'-shield and r is incident with no vertex except x and y. We deduce that $k' = k$, as required. □

Consequently, we have the following.

(3.4). *Let G be a graph in a sphere Σ and let* big *be a majority of order $k \geq 0$. For any noose $F \subseteq \Sigma$ of length $\leq k$, $|V(G) \cap \text{big}(F)| \geq \frac{1}{6}k^2 + \frac{5}{6}k + \frac{2}{3}$.*

Proof. From the definition of optimal noose, there is an optimal noose F' with $G \cap \text{big}(F') \subseteq G \cap \text{big}(F)$. We claim that $|V(G \cap \text{big}(F'))| \geq \frac{1}{6}k^2 + \frac{5}{6}k + \frac{2}{3}$. If

$k \leq 2$, this follows from the second axiom (together with the first if $k = 2$), while, for $k \geq 3$, it follows from (2.5), since $G \cap \text{big}(F')$ is a strong k-shield by (3.3). Since $|V(G \cap \text{big}(F))| \geq |V(G \cap \text{big}(F'))|$, the result follows. □

Let us say a noose in G has *discrepancy* $|n_1 - n_2|$, where it bounds closed discs Δ_1, Δ_2, and $n_i = |V(G) \cap \Delta_i|$ $(i = 1, 2)$. We have immediately from (3.4) the following.

(3.5). *Let G be a graph in a sphere Σ, with n vertices. There is a noose of length $\leq 6^{1/2}n^{1/2}$ with discrepancy $\leq \frac{1}{3}n$.*

Proof. Let $k = \lfloor 6^{1/2}n^{1/2} \rfloor$. If G has a majority of order k then, by (3.4),

$$|V(G)| \geq \tfrac{1}{6}k^2 + \tfrac{5}{6}k + \tfrac{2}{3} \geq \tfrac{1}{6}(k+1)^2 > n,$$

a contradiction. Thus, by (3.1) (with $w(v) = 1$ for all v), there is a noose F of length $\leq k$ such that the discs D_1, D_2 bounded by F satisfy

$$|(D_i - F) \cap V(G)| + \tfrac{1}{2}|F \cap V(G)| \leq \tfrac{2}{3}|V(G)| \qquad (i = 1, 2),$$

or, equivalently, that F has discrepancy $\leq \frac{1}{3}n$. □

Actually, here the $6^{1/2}$ is irrelevant; all we need from (3.5) is that some noose has discrepancy $\leq \frac{1}{2}n$.

(3.6). *Let G be a graph in a sphere Σ, with n vertices. There is a noose of length $\leq \frac{3}{2}(2n)^{1/2}$ with discrepancy $\leq \frac{1}{2}n$.*

Proof. We assume that $n > 0$. By (3.5), there is a noose with discrepancy $\leq \frac{1}{2}n$. Let us choose such a noose F of minimum order, say k. Let F bound closed discs Δ, Δ', with $|V(G) \cap \Delta| \geq |V(G) \cap \Delta'|$.

(1). *$G \cap \Delta$ is a k-shield in Δ.*

For let $F_3 \subseteq \Delta$ be a G-normal I-arc with ends $x, y \in F$ and with $F \cap bd(\Delta) = \{x, y\}$, and let F_1, F_2 be the two I-arcs in F with ends x, y. Suppose that

$$|F_3 \cap V(G)| < |F_1 \cap V(G)|, \ |F_2 \cap V(G)|.$$

Since

$$|V(G) \cap (\Delta - F)| + \tfrac{1}{2}|V(G) \cap F| \geq \tfrac{1}{2}n$$

because $|V(G) \cap \Delta| \geq |V(G) \cap \Delta'|$, we may assume that

$$|V(G) \cap (\Delta_1 - (F_1 \cup F_3))| + \tfrac{1}{2}|V(G) \cap (F_1 \cup F_3)| \geq \tfrac{1}{4}n$$

without loss of generality, where Δ_1 is the closed disc in Δ bounded by $F_1 \cup F_3$. But then, $F_1 \cup F_3$ has discrepancy $\leq \frac{1}{2}n$ and has order $< k$, contrary to the choice of F. This proves (1), by (2.4).

Now let us choose such F, Δ with $E(G \cap \Delta)$ minimal. It follows that $G \cap \Delta$ is a stable k-shield, and so, by (2.3),

$$|V(G \cap \Delta)| \geq \tfrac{1}{6}k^2 + \tfrac{1}{2}k + 1,$$

(for we may assume that $k > \frac{3}{2}(2n)^{1/2}$, since otherwise F satisfies the theorem, and $\frac{3}{2}(2n)^{1/2} \geq 2$, since $n \geq 1$; so $k \geq 3$). Hence,

$$|V(G \cap \Delta') \geq \tfrac{1}{6}k^2 + \tfrac{1}{2}k + 1 - \tfrac{1}{2}n,$$

since F has discrepancy $\leq \frac{1}{2}n$; so

$$\begin{aligned} n + k = |V(G)| + |V(G) \cap F| &= |V(G \cap \Delta)| + |V(G \cap \Delta')| \\ &\geq 2(\tfrac{1}{6}k^2 + \tfrac{1}{2}k + 1) - \tfrac{1}{2}n. \end{aligned}$$

It follows that $\frac{3}{2}n \geq \frac{1}{3}k^2 + 2$, and so $k < \frac{3}{2}(2n)^{1/2}$, as required. □

We deduce the following.

(3.7). *Let G be a graph in a sphere Σ, with n vertices and with a majority of order k. Then $k \leq \frac{3}{2}(2n)^{1/2} - 1$.*

Proof. Let big be a majority of order k and suppose that $k \geq \lfloor\frac{3}{2}(2n)^{1/2}\rfloor$. By (3.6), there is a noose F of length $\leq k$ with discrepancy $\leq \frac{1}{2}n$. By (3.4),

$$|V(G) \cap \mathrm{big}(F)| \geq \tfrac{1}{6}k^2 + \tfrac{5}{6}k + \tfrac{2}{3};$$

but

$$|V(G)| + |V(G) \cap F| \geq 2|V(G) \cap \mathrm{big}(F)| - \tfrac{1}{2}n,$$

since F has discrepancy $\leq \frac{1}{2}n$, and so

$$\tfrac{3}{2}n + k \geq \tfrac{1}{3}k^2 + \tfrac{5}{3}k + \tfrac{4}{3},$$

since $|V(G) \cap F| \leq k$. Hence, $\frac{1}{3}(k+1)^2 + 1 \leq \frac{3}{2}n$, and so $k + 1 < \frac{3}{2}(2n)^{1/2}$, a contradiction. Thus $k < \lfloor\frac{3}{2}(2n)^{1/2}\rfloor$, as required. □

From (3.7) and (3.1), we deduce our main result, the following weighted version of (2.1).

(3.8). *Let G be a graph in a sphere with n vertices and, for each vertex v, let $w(v) \geq 0$ be a real number. There is a noose F with $|F \cap V(G)| \leq \frac{3}{2}(2n)^{1/2}$ such that*

$$w((D - F) \cap V(G)) + \tfrac{1}{2}w(F \cap V(G)) \leq \tfrac{2}{3}w(V(G))$$

for both closed discs D bounded by F.

Proof. Let $k = \lfloor\frac{3}{2}(2n)^{1/2}\rfloor$. By (3.7), G has no majority of order k, and the result follows from (3.1). □

Similarly, we can use (3.7) and (3.2) to deduce a $\frac{3}{2}(2n)^{1/2}$-separator result when the weights are on the edges.

Finally, let us show the following curiosity, which indicates that, for finding "separating" nooses of length $\leq \lambda n^{1/2}$ where $\lambda \geq 2$, in some sense the unweighted case is the hardest.

(3.9). *Let G be a graph in a sphere with n vertices, let $k \geq 2n^{1/2} - 1$ be an integer, and suppose that there is a noose F^* of length $\leq k$ such that*

$$|(D - F^*) \cap V(G)| + \tfrac{1}{2}|F^* \cap V(G)| \leq \tfrac{2}{3}|V(G)|$$

for both closed discs D bounded by F^*. Then

(i) *G has no majority of order k;*

(ii) *for any function $w : V(G) \to \mathbf{R}_+$, there is a noose F of length $\leq k$ such that*

$$w((D - F) \cap V(G)) + \tfrac{1}{2}w(F \cap V(G)) \leq \tfrac{2}{3}w(V(G))$$

for both closed discs D bounded by F;

(iii) *for any function $w : E(G) \to \mathbf{R}_+$ such that $w(f) \leq \frac{2}{3}w(E(G))$ for every $f \in E(G)$, there is a noose F of length $\leq k$ such that $w(E(G \cap D)) \leq \frac{2}{3}w(E(G))$ for both closed discs D bounded by F.*

Proof. Suppose that big is a majority of order k. By (3.4),

$$|V(G) \cap \mathrm{big}(F^*)| \geq \tfrac{1}{6}k^2 + \tfrac{5}{6}k + \tfrac{2}{3}.$$

By the hypothesis,

$$|V(G) \cap \mathrm{big}(F^*)| - \tfrac{1}{2}|V(G) \cap F^*| \leq \tfrac{2}{3}n.$$

Moreover, $|V(G) \cap F^*| \leq k$, and so

$$\tfrac{1}{6}k^2 + \tfrac{5}{6}k + \tfrac{2}{3} - \tfrac{1}{2}k \leq \tfrac{2}{3}n;$$

that is, $(k+1)^2+3 \leq 4n$. But $k+1 \geq 2n^{1/2}$, a contradiction. This proves (i), and (ii) and (iii) follow from (3.1) and (3.2), respectively. □

REFERENCES

[1] H. Djidjev, *On the problem of partitioning planar graphs*, SIAM J. Algebraic Discrete Meth., 3 (1982), pp. 229–240.

[2] H. Gazit, *An algorithm for finding a $\frac{7}{3}\cdot\sqrt{n}$ separator in planar graphs*, submitted.

[3] R. J. Lipton and R. E. Tarjan, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–189.

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

[5] S. Randby, *Embedding K_5 in 4-Connected Graphs*, Ph.D. thesis, Dept. of Math., Ohio State University, Columbus, OH, 1991.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 194-202, May 1994

A STOCHASTIC PROCESS INTERPRETATION OF PARTITION IDENTITIES*

G. M. CONSTANTINE† AND T. H. SAVITS‡

Abstract. The notion of compound nonhomogeneous Poisson processes provides a general framework for Bell polynomials and set partitions. Identities that can be viewed as generalizations of Dobinski's formula are obtained by interpreting them as moments of such processes.

Key words. nonhomogeneous Poisson process, compound Poisson process, independent increments, Bell polynomials, set partitions, inversion formula, Dobinski identity

AMS subject classifications. primary 05A19, secondary 60G07.

1. Introduction. One of the classical identities in combinatorial theory concerns the Bell numbers and is due to Dobinski. The Bell number B_n counts the number of partitions of a set with n elements. Rota (1964) describes many problems of enumeration related to the lattice of partitions of a finite set. An explicit formula for B_n, written as a finite sum in terms of set partitions, is

$$B_n = \sum_{r=1}^{n} \sum_{\substack{\gamma_1+\cdots+\gamma_n=r \\ \gamma_1+2\gamma_2+\cdots+n\gamma_n=n}} \frac{n!}{(1!)^{\gamma_1}\cdots(n!)^{\gamma_n}(\gamma_1!)\cdots(\gamma_n!)}. \tag{1.1}$$

The exponential generating function for B_n coincides with the moment generating function of a Poisson random variable with mean 1, i.e.,

$$\sum_{n=0}^{\infty} B_n \frac{x^n}{n!} = \exp\{e^x - 1\}. \tag{1.2}$$

A remarkable formula of Dobinski (1877) expresses B_n as a fast converging series, namely,

$$B_n = e^{-1} \sum_{m=1}^{\infty} \frac{m^n}{m!}. \tag{1.3}$$

This formula is usually established using generating functions as in, for example, Constantine (1987). A different proof is given in Rota (1964).

Recently, we were led to another proof based on Poisson processes. In fact, substantial generalizations of (1.3) arise by interpreting the two sides as moments of certain stochastic processes.

In §2 we introduce the necessary background material on stochastic processes and derive a general identity. Special cases of the identity appear in §3. A general pair of inverse relations is established in §4. The last section presents some further results including convolution identities for Bell numbers.

Since summations over set partitions arise frequently in this work, we denote by $\sum_{\Gamma(r,n)}$ the sum over all indices in

$$\Gamma(r, n) = \left\{(\gamma_1, \ldots, \gamma_n) : \gamma_i \text{ nonnegative integers}, \sum_{i=1}^{n} \gamma_i = r, \sum_{i=1}^{n} i\gamma_i = n\right\}.$$

* Received by the editors June 17, 1992; accepted for publication (in revised form) March 2, 1993.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260. This author was supported by a Fulbright grant.

‡ Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260. This author was supported by National Security Agency grant MDA-904-90-H-4036.

Thus, for example, (1.1) becomes

$$B_n = \sum_{r=1}^{n} \sum_{\Gamma(r,n)} n! \prod_{j=1}^{n} \frac{1}{(j!)^{\gamma_j}(\gamma_j!)}.$$

2. A general identity. Our main result involves the notion of compound nonhomogeneous Poisson processes. Thus we first review some basic definitions. See also Ross (1983) for a general exposition on stochastic processes.

A stochastic process $N = \{N(t), t \geq 0\}$ is called a counting process if, with probability 1, its sample paths are nonnegative right-continuous step-functions starting at 0 and only increasing by jumps of size 1. A stochastic process $X = \{X(t),\ t \geq 0\}$ is said to have independent increments if, for all $n = 1, 2, \ldots$ and every choice $0 \leq t_0 < t_1 < \cdots < t_n$, the random variables $X(t_1) - X(t_0), \ldots, X(t_n) - X(t_{n-1})$ are independent. A counting process having independent increments is called a nonhomogeneous Poisson process.

If $N = \{N(t),\ t \geq 0\}$ is a nonhomogeneous Poisson process, we denote its mean function $E[N(t)]$ by $\Lambda(t)$. If $\Lambda(t) = \lambda t$ for some $\lambda > 0$, then N is called a Poisson process with rate λ. In this case, N also has stationary increments, i.e., the distribution of $N(t + s) - N(t)$ is independent of t; moreover, its distribution is Poisson with parameter λs.

We now consider a nonhomogeneous Poisson process $N = \{N(t), t \geq 0\}$ with mean function $\Lambda(t)$. *In what follows, we always assume that* $\Lambda(t)$ *is a continuous function of t.* Clearly, $\Lambda(t)$ must also be increasing since $N(t)$ increases. Since the process only increases by jumps of size 1, it is completely characterized by its arrival (or jump) times S_1, S_2, ... We now attach a "cost" Y_k to the kth arrival time and denote the corresponding "cost" process by $X = \{X(t), t \geq 0\}$ as follows:

$$X(t) = \sum_{k=1}^{N(t)} Y_k. \tag{2.1}$$

More precisely, the process X is prescribed by requiring that the random variables Y_1, $Y_2, \ldots$ be conditionally independent, given $S_1, S_2, \ldots$, in the sense that, for every $n = 1, 2, \ldots$ and all Borel sets $B_1, B_2, \ldots, B_n$ of the real line $\mathbf{R}$, we have

$$P\{Y_1 \in B_1, \ldots, Y_n \in B_n | S_1, \ldots, S_n\} = \prod_{k=1}^{n} L(S_k; B_k), \tag{2.2}$$

where $L(s; B)$ is a stochastic kernel on $[0, \infty) \times \mathbf{R}$. This means that, for every Borel set B, $L(s; B)$ is Borel-measurable in s, and, for every fixed $s \geq 0$, $L(s; dy)$ is a probability measure on the Borel subsets of $\mathbf{R}$.

Such processes were investigated in Chen and Savits (1990), (1993) and were called compound nonhomogeneous Poisson processes with representation (L, Λ). In particular, it was shown that X is an independent increment process with characteristic function

$$\phi(u; t) = E[e^{iuX(t)}] = \exp\left\{\int_0^t \int_{\mathbf{R}} (e^{iuy} - 1)L(s; dy)\, d\Lambda(s)\right\}. \tag{2.3}$$

Our general Dobinski-like identity results from consideration of the moments of the above process. Suppose that

$$\int_0^t \int_{\mathbf{R}} |y|^n L(s; dy)\, d\Lambda(s) < \infty$$

for some n and $t > 0$ (and, consequently, for all $0 \leq m \leq n$ and $0 \leq s \leq t$). Then it can be shown (cf. §2 of Chen and Savits (1990)) that $X(t)$ has a finite moment of order n. In this case, a direct calculation yields

$$\begin{aligned} E[X^n(t)] &= E\left[\left(\sum_{k=1}^{N(t)} Y_k\right)^n\right] \\ &= \sum_{m=1}^{\infty} E\left[\left(\sum_{k=1}^{m} Y_k\right)^n ; N(t) = m\right] \\ &= \sum_{m=1}^{\infty} E\left[\sum_{\alpha_1 + \cdots + \alpha_m = n} \binom{n}{\alpha_1 \cdots \alpha_m} \prod_{j=1}^{m} Y_j^{\alpha_j}; S_m \le t < S_{m+1}\right] \\ &= e^{-\Lambda(t)} \sum_{m=1}^{\infty} \int_0^t \int_{s_1}^t \cdots \int_{s_m}^t \left\{ \sum_{\alpha_1 + \cdots + \alpha_m = n} \binom{n}{\alpha_1 \cdots \alpha_m} \right. \\ &\qquad \left. \times \left[\prod_{j=1}^{m} \int_{\mathbf{R}} y_j^{\alpha_j} L(s_j; dy_j)\right] d\Lambda(s_m) \cdots d\Lambda(s_1)\right\} \\ &= e^{-\Lambda(t)} \sum_{m=1}^{\infty} \frac{1}{m!} \sum_{\alpha_1 + \cdots + \alpha_m = n} \binom{n}{\alpha_1 \cdots \alpha_m} \prod_{j=1}^{m} \mu_{\alpha_j}(t), \end{aligned} \tag{2.4}$$

where

$$\mu_j(t) = \int_0^t \int_{\mathbf{R}} y^j L(s; dy)\, d\Lambda(s).$$

In the above, the second equality follows, since the event $N(t) = m$ is equivalent to the event $S_m \le t < T_{m+1}$; the third equality follows, since the joint distribution of $(S_1, \ldots, S_m, S_{m+1})$ is $e^{-\Lambda(s_{m+1})} d\Lambda(s_{m+1}) d\Lambda(s_m) \cdots d\Lambda(s_1)$ on $0 \le s_1 < \cdots < s_m < s_{m+1}$; the last equality uses the fact that the integrand is permutation invariant.

On the other hand, the nth moment of $X(t)$ can also be calculated from its characteristic function given in (2.3) as

$$E[X^n(t)] = i^{-n} \frac{\partial^n}{\partial u^n} \phi(u; t)|_{u=0}.$$

According to Faa Di Bruno's formula (e.g., see Constantine (1987) or Comtet (1974)) for computing higher-order derivatives of a composition of functions, we obtain that

$$i^{-n} \frac{\partial^n}{\partial u^n} \phi(u; t)|_{u=0} = \sum_{r=1}^{n} \sum_{\Gamma(r,n)} \frac{n!}{(1!)^{\gamma_1} \cdots (n!)^{\gamma_n} (\gamma_1 !) \cdots (\gamma_n !)} \prod_{j=1}^{n} \mu_j^{\gamma_j}(t). \tag{2.5}$$

Equating (2.4) and (2.5) yields the main result, below.

THEOREM 2.1. *Let* $\Lambda(t)$ *be a continuous increasing function on* $[0, \infty)$ *and* $L(s; dy)$ *a stochastic kernel on* $[0, \infty) \times \mathbf{R}$ *and suppose that* $\int_0^t \int_{\mathbf{R}} |y|^n L(s; dy)\, d\Lambda(s) < \infty$. *Then*

$$e^{-\Lambda(t)} \sum_{m=1}^{\infty} \frac{1}{m!} \sum_{\alpha + \cdots + \alpha_m = n} \prod_{j=1}^{m} \frac{\mu_{\alpha_j}(t)}{\alpha_j !} = \sum_{r=1}^{n} \sum_{\Gamma(r,n)} \prod_{j=1}^{n} \frac{\mu_j^{\gamma_j}(t)}{(j!)^{\gamma_j} (\gamma_j !)}, \tag{2.6}$$

where $\mu_j(t) = \int_0^t \int_{\mathbf{R}} y^j L(s; dy)\, d\Lambda(s)$.

In particular, if we take $L(s; dy) = \delta_{\{1\}}(dy)$ to be the Dirac measure at $y = 1$ and choose $\Lambda(s) = s$, then the process X reduces to the Poisson process with rate $\lambda = 1$. In this case, $\mu_j(t) = t$ for all j, and Dobinski's formula (1.3) follows from the corollary below when we set $t = 1$.

COROLLARY. *Let $\Lambda(s) = s$ and $L(s; dy) = \delta_{\{1\}}(dy)$. Then*

$$e^{-t}\sum_{m=1}^{\infty}\frac{t^m m^n}{m!}=\sum_{r=1}^{n}t^r\sum_{\Gamma(r,n)}\frac{n!}{(1!)^{\gamma_1}\cdots(n!)^{\gamma_n}(\gamma_1!)\cdots(\gamma_n!)}. \tag{2.7}$$

Remark 2.1. The above can also be written as

$$e^{-t}\sum_{m=1}^{\infty}\frac{t^m m^n}{m!}=\sum_{r=1}^{n}t^r S_n^r,$$

where

$$S_n^r=\sum_{\Gamma(r,n)}\frac{n!}{(1!)^{\gamma_1}\cdots(n!)^{\gamma_n}(\gamma_1!)\cdots(\gamma_n!)} \tag{2.8}$$

is the Stirling number of the second kind; it counts the number of partitions of a set with n elements into exactly r classes.

If we assume that $X(t)$ has finite moments of all orders, then, by a formal power series expansion, we obtain

$$\exp\left\{\int_0^t\int_{\mathbf{R}}(e^{iuy}-1)L(s;dy)\,d\Lambda(s)\right\}=E[e^{iuX(t)}]=\sum_{n=0}^{\infty}\frac{(iu)^n}{n!}E[X^n(t)]. \tag{2.9}$$

This can be viewed as a (formal) generating function identity for the moments of $X(t)$. Of course, under appropriate summability conditions, the formal identity becomes a true identity. In particular, when $L(s; dy) = \delta_{\{1\}}(dy)$ and $\Lambda(s) = s$, $E[X^n(1)] = B_n$, and so (2.9) yields the following characteristic function version of (1.2):

$$\exp\{e^{iu}-1\}=\sum_{n=0}^{\infty}\frac{(iu)^n}{n!}B_n. \tag{2.10}$$

3. Examples.

I. For this class of examples, we let $L(s; dy) = L(dy)$ be independent of s. This means that the costs $Y_1, Y_2, \ldots$ in (2.1) are independent and identically distributed random variables with distribution function L. Let

$$m_j=\int_{\mathbf{R}}y^j\,dL(y)$$

denote its jth moment. Thus $\mu_j(t) = m_j\Lambda(t)$, and so (2.6) becomes

$$e^{-\Lambda(t)}\sum_{k=1}^{\infty}\frac{\Lambda^k(t)}{k!}\sum_{\alpha_1+\cdots+\alpha_k=n}\prod_{j=1}^{k}\frac{m_{\alpha_j}}{\alpha_j!}=\sum_{r=1}^{n}\Lambda^r(t)\sum_{\Gamma(r,n)}\prod_{j=1}^{n}\frac{1}{\gamma_j!}\left(\frac{m_j}{j!}\right)^{\gamma_j}. \tag{3.1}$$

Some special subcases follow, all with $\Lambda(t) = t$.

Remark 3.1. When all the odd moments m_{2j+1} are zero, identity (3.1) can be rewritten as

$$e^{-\Lambda(t)}\sum_{k=1}^{\infty}\frac{\Lambda^k(t)}{k!}\sum_{\alpha_1+\cdots+\alpha_k=n}\prod_{j=1}^{k}\frac{m_{2\alpha_j}}{(2\alpha_j)!}=\sum_{r=1}^{n}\Lambda^r(t)\sum_{\Gamma(r,n)}\prod_{j=1}^{n}\frac{1}{\gamma_j!}\left(\frac{m_{2j}}{(2j)!}\right)^{\gamma_j}.$$

This form will be used in deriving identities (3.2) and (3.5). An analogous version of (2.6) will also be used in deriving (3.9).

(i) Uniform distribution on $[-1, 1]$:

$$m_j = \frac{1}{2}\int_{-1}^{1} y^j\,dy = \begin{cases} \dfrac{1}{j+1} & \text{if } j \text{ is even,} \\ 0 & \text{if } j \text{ is odd.} \end{cases}$$

Thus (3.1) becomes

$$e^{-t}\sum_{k=1}^{\infty}\frac{t^k}{k!}\sum_{\alpha_1+\cdots+\alpha_k=n}\prod_{j=1}^{k}\frac{1}{(2\alpha_j+1)!} = \sum_{r=1}^{n} t^r \sum_{\Gamma(r,n)}\prod_{j=1}^{n}\frac{1}{[(2j+1)!]^{\gamma_j}(\gamma_j!)}\,. \tag{3.2}$$

(ii) Exponential distribution:

$$m_j = \int_0^{\infty} y^j e^{-y}\,dy = j!, \quad \text{and thus}$$

$$e^{-t}\sum_{k=1}^{\infty}\binom{k+n-1}{n}\frac{t^k}{k!} = \sum_{r=1}^{n} t^r \sum_{\Gamma(r,n)}\prod_{j=1}^{n}\frac{1}{(\gamma_j!)}\,. \tag{3.3}$$

(iii) Gamma distribution:

$$m_j = \frac{1}{\Gamma(\alpha)}\int_0^{\infty} y^j y^{\alpha-1} e^{-y}\,dy = \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)} \quad \text{for } \alpha > 0\text{, which yields}$$

$$\begin{aligned} &e^{-t}\sum_{k=1}^{\infty}\frac{1}{k!}\left[\frac{t}{\Gamma(\alpha)}\right]^k \sum_{\alpha_1+\cdots+\alpha_k=n}\prod_{j=1}^{k}\frac{\Gamma(\alpha+\alpha_j)}{\Gamma(1+\alpha_j)} \\ &= \sum_{r=1}^{n}\left[\frac{t}{\Gamma(\alpha)}\right]^r \sum_{\Gamma(r,n)}\prod_{j=1}^{n}\frac{[\Gamma(\alpha+j)/\Gamma(1+j)]^{\gamma_j}}{\Gamma(1+\gamma_j)}\,. \end{aligned} \tag{3.4}$$

(iv) Normal distribution:

$$m_j = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^j e^{-y^2/2}\,dy = \begin{cases} \dfrac{j!}{(j/2)!2^{j/2}} & \text{if } j \text{ is even,} \\ 0 & \text{if } j \text{ is odd.} \end{cases}$$

We then obtain

$$e^{-t}\sum_{k=1}^{\infty}\frac{t^k k^n}{k!n!} = \sum_{r=1}^{n} t^r \sum_{\Gamma(r,n)}\sum_{j=1}^{n}\frac{1}{(j!)^{\gamma_j}(\gamma_j!)}\,. \tag{3.5}$$

This identity is the same as Dobinski's identity (2.7).

(v) Poisson distribution:

Since

$$m_j = e^{-1}\sum_{k=1}^{\infty}\frac{k^j}{k!} = B_j,$$

we obtain the following identity involving Bell numbers:

$$e^{-t}\sum_{k=1}^{\infty}\frac{t^k}{k!}\sum_{\alpha_1+\cdots+\alpha_k=n}\prod_{j=1}^{k}\frac{B_{\alpha_j}}{(\alpha_j!)} = \sum_{r=1}^{n} t^r \sum_{\Gamma(r,n)}\prod_{j=1}^{n}\frac{(B_j/j!)^{\gamma_j}}{\gamma_j!}\,. \tag{3.6}$$

II. For this set of examples, we consider a parametric family of distributions, say $\{F(y; \beta), \beta \in B\}$. Let $b : [0, \infty) \to B$ be Borel measurable and set $L(s; dy) = F(dy; b(s))$. Since L now depends on s, identity (3.1) is no longer applicable; instead we use (2.6).

We illustrate with some particular examples for the case where $B = [0, \infty)$ and $b(s) = \Lambda(s)$.

(i) $F(y; \beta)$ uniform on $[0, \beta]$:

$$\mu_j(t) = \int_0^t \int_{\mathbf{R}} y^j L(s; dy)\, d\Lambda(s) = \int_0^t \frac{1}{\Lambda(s)} \int_0^{\Lambda(s)} y^j\, dy\, d\Lambda(s) = \frac{1}{(j+1)^2} \Lambda^{j+1}(t).$$

Hence, we have the identity

$$(3.7)\quad e^{-\Lambda(t)} \sum_{k=1}^{\infty} \frac{\Lambda^k(t)}{k!} \sum_{\alpha_1+\cdots+\alpha_k=n} \prod_{j=1}^{k} \frac{1}{(\alpha_j+1)^2(\alpha_j!)} = \sum_{r=1}^{n} \Lambda^r(t) \sum_{\Gamma(r,n)} \frac{1}{[(j+1)^2 j!]^{\gamma_j}(\gamma_j!)}\,.$$

(ii) $F(y; \beta)$ gamma with scale parameter β (and shape parameter α):

$$\mu_j(t) = \int_0^t \frac{1}{\Lambda^{\alpha}(s)\Gamma(\alpha)} \int_0^{\infty} y^j y^{\alpha-1} e^{-y/\Lambda(s)}\, dy\, d\Lambda(s) = \frac{\Lambda^{j+1}(t)\Gamma(\alpha+j)}{(j+1)\Gamma(\alpha)}\,.$$

Consequently,

$$(3.8)\quad \begin{aligned} &e^{-\Lambda(t)} \sum_{k=1}^{\infty} \frac{1}{k!} \left[\frac{\Lambda(t)}{\Gamma(\alpha)}\right]^k \sum_{\alpha_1+\cdots+\alpha_k=n} \prod_{j=1}^{k} \frac{\Gamma(\alpha+\alpha_j)}{\Gamma(2+\alpha_j)} \\ &= \sum_{r=1}^{n} \left[\frac{\Lambda(t)}{\Gamma(\alpha)}\right]^r \sum_{\Gamma(r,n)} \prod_{j=1}^{n} \frac{[\Gamma(\alpha+j)/\Gamma(2+j)]^{\gamma_j}}{\Gamma(1+\gamma_j)}\,. \end{aligned}$$

(iii) $F(y; \beta)$ normal with variance β (and mean zero):

$$\mu_j(t) = \int_0^t \frac{1}{\sqrt{2\pi\Lambda(s)}} \int_{\mathbf{R}} y^j e^{-y^2/2\Lambda(s)}\, dy\, d\Lambda(s) = \begin{cases} \dfrac{j!\Lambda^{(j/2)+1}(t)}{(j/2)!2^{(j/2)}[(j/2)+1]} & \text{if } j \text{ is even,} \\ 0 & \text{if } j \text{ is odd.} \end{cases}$$

Thus,

$$(3.9)\quad e^{-\Lambda(t)} \sum_{k=1}^{\infty} \frac{\Lambda^k(t)}{k!} \sum_{\alpha_1+\cdots+\alpha_k=n} \prod_{j=1}^{k} \frac{1}{(\alpha_j+1)!} = \sum_{r=1}^{n} \Lambda^r(t) \sum_{\Gamma(r,n)} \prod_{j=1}^{n} \frac{1}{[(j+1)!]^{\gamma_j}\gamma_j!}\,.$$

4. A pair of inverse relations. Let f be a complex-valued function defined on $\{1, 2, \ldots, n\}$. It can then be shown that there exists a probability measure L on the complex plane such that

$$\int_{\mathbf{C}} y^j L(dy) = j! f(j)$$

for all $j = 1, 2, \ldots, n$. In fact, we can choose $x_1, \ldots, x_n \in \mathbf{C}$ such that $L = (1/n) \sum_{i=1}^{n} \delta_{\{x_i\}}$, i.e., L is the discrete uniform probability on $\{x_1, \ldots, x_n\}$. It then follows from (3.1), with $\Lambda(t) = t$, that

$$(4.1)\quad e^{-t} \sum_{k=1}^{\infty} \frac{t^k}{k!} \sum_{\alpha_1+\cdots+\alpha_k=n} \prod_{j=1}^{k} f(\alpha_j) = \sum_{r=1}^{n} t^r \sum_{\Gamma(r,n)} \prod_{j=1}^{n} \frac{f(j)^{\gamma_j}}{\gamma_j!}\,.$$

We now define, for $n \geq 1$ and $k \geq 1$,

$$S(f, k, n) = \sum_{\alpha_1 + \cdots + \alpha_k = n} \prod_{j=1}^{k} f(\alpha_j) \tag{4.2}$$

and

$$T(f, k, n) = \sum_{\Gamma(k,n)} \prod_{j=1}^{n} \frac{f(j)^{\gamma_j}}{\gamma_j!}. \tag{4.3}$$

We also set $S(f, 0, n) = T(f, 0, n) = 0$ and note that $T(f, k, n) = 0$ whenever $k > n$. Thus (4.1) takes the form

$$e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} S(f, k, n) = \sum_{r=0}^{n} t^r T(f, r, n). \tag{4.4}$$

The following identities result by equating like powers of t.

THEOREM 4.1. *Let $f(1), \ldots, f(n)$ be n complex numbers and define $S(f, k, n)$, $T(f, k, n)$ as in* (4.2), (4.3), *respectively. Then*

$$e^{-t} \sum_{k=1}^{\infty} \frac{t^k}{k!} S(f, k, n) = \sum_{r=0}^{n} t^r T(f, r, n),$$

$$S(f, k, n) = \sum_{r=0}^{k} \binom{k}{r} r! T(f, r, n),$$

and

$$T(f, k, n) = \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} \frac{1}{k!} S(f, r, n)$$

for all $k \geq 0$.

Special cases.

(i) $f(j) = 1$, $1 \leq j \leq n$:

Then

$$S(f, k, n) = \binom{n+k-1}{n} \quad \text{and} \quad T(f, k, n) = \sum_{\Gamma(k,n)} \prod_{j=1}^{k} \frac{1}{\gamma_j!}.$$

Thus we obtain the identity

$$k! \sum_{\Gamma(k,n)} \prod_{j=1}^{k} [\gamma_j!]^{-1} = \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} \binom{n+r-1}{n},$$

which expresses $T(1, k, n)$ in terms of binomial numbers only.

(ii) $f(j) = 1/j!$, $1 \leq j \leq n$:

$$S(f, k, n) = \sum_{\alpha_1 + \cdots + \alpha_k = n} \prod_{j=1}^{k} [\alpha_j!]^{-1} = \frac{k^n}{n!}$$

and

$$T(f, k, n) = \sum_{\Gamma(k,n)} \prod_{j=1}^{n} [\gamma_j!(j!)^{\gamma_j}]^{-1} = \frac{1}{n!} S_n^k,$$

where S_n^k is the Stirling number of the second kind (see (2.8)). Hence Theorem 4.1 yields the well-known identity

$$(k!)S_n^k = \sum_{r=0}^{k} (-1)^{k-r}\binom{k}{r} r^n,$$

which expresses the Stirling numbers in terms of binomial coefficients.

(iii) $f(j) = j$, $1 \leq j \leq n$:

$$S(f, k, n) = \sum_{\alpha_1 + \cdots + \alpha_k = n} \prod_{j=1}^{k} \alpha_j,$$

$$T(f, k, n) = \sum_{\Gamma(k,n)} \prod_{j=1}^{n} \frac{j^{\gamma_j}}{\gamma_j!},$$

and so,

$$e^{-t} \sum_{k=1}^{\infty} \frac{t^k}{k!} \sum_{\alpha_1 + \cdots + \alpha_k = n} \prod_{j=1}^{k} \alpha_j = \sum_{r=0}^{n} t^r \sum_{\Gamma(r,n)} \prod_{j=1}^{n} \frac{j^{\gamma_j}}{\gamma_j!}.$$

5. Other results.

I. *Transformations.* Let X be a compound nonhomogeneous Poisson process with representation (L, Λ) as in (2.1) and let $\phi : [0, \infty) \times \mathbf{R} \to \mathbf{R}$ be Borel-measurable. Then $L^{\phi}(s; B) = L(s; \phi_s^{-1}[B])$ is also a stochastic kernel on $[0, \infty) \times \mathbf{R}$, where we define $\phi_s(y) = \phi(s, y)$. Hence there exists a compound nonhomogeneous Poisson process X^{ϕ} with representation (L^{ϕ}, Λ). Furthermore,

$$\mu_j^{\phi}(t) = \int_0^t \int_{\mathbf{R}} y^j L^{\phi}(s; dy)\, d\Lambda(s) = \int_0^t \int_{\mathbf{R}} \phi^j(s, y) L(s; dy)\, d\Lambda(s). \tag{5.1}$$

Example. Let $\phi(s, y) = y^2$, $L(s; dy) = L(dy) = (\sqrt{2\pi})^{-1} e^{-y^2/2} dy$ be the normal distribution, and $\Lambda(t) = t$. Then

$$\mu_j^{\phi}(t) = \frac{(2j)!}{j! 2^j}\, t,$$

and therefore we obtain the identity

$$e^{-t} \sum_{k=1}^{\infty} \frac{t^k}{k!} \sum_{\alpha_1 + \cdots + \alpha_k = n} \prod_{j=1}^{k} \frac{(2\alpha_j)!}{\alpha_j! \alpha_j!} = \sum_{r=1}^{n} t^r \sum_{\Gamma(r,n)} \prod_{j=1}^{n} \left(\frac{(2j)!}{j! j!}\right)^{\gamma_j} \frac{1}{\gamma_j!}. \tag{5.2}$$

II. *Independent increments.* It was noted in (2.3) that, if X is a compound nonhomogeneous Poisson process, then X has independent increments (see §2 for the definition). Thus, for all $0 \leq s < t$, we can write

$$X(t) = X(s) + [X(t) - X(s)],$$

where $X(s)$ and $[X(t) - X(s)]$ are independent. Hence

$$E[X^n(t)] = \sum_{k=0}^{n} \binom{n}{k} E[X^k(s)] E[\{X(t) - X(s)\}^{n-k}]. \tag{5.3}$$

In particular, if X is a Poisson process with rate $\lambda = 1$ (i.e., $L(s; dy) = \delta_{\{1\}}(dy)$ and $\Lambda(s) = s$), then the increments are Poisson random variables: $X(s)$ is Poisson with parameter s and $[X(t) - X(s)]$ is Poisson with parameter $(t - s)$. Therefore, for $s = 1$ and

$t = 2$,

$$E[X^n(2)] = \sum_{k=0}^{n} \binom{n}{k} E[X^k(1)]E[X^{n-k}(1)].$$

However, $E[X^k(1)] = B_k$, the kth Bell number. Also, from (2.7),

$$E[X^n(2)] = e^{-2} \sum_{k=1}^{\infty} \frac{2^k k^n}{k!} = \sum_{r=1}^{n} 2^r S_n^r.$$

Hence we obtain

$$\sum_{k=0}^{n} \binom{n}{k} B_k B_{n-k} = e^{-2} \sum_{k=1}^{\infty} \frac{2^k k^n}{k!} = \sum_{k=1}^{n} 2^r S_n^r. \tag{5.4}$$

More generally, writing

$$X(q) = \sum_{l=1}^{q} [X(l) - X(l-1)],$$

we obtain the identity

$$\sum_{\alpha_1 + \cdots + \alpha_q = n} \binom{n}{\alpha_1 \cdots \alpha_q} B_{\alpha_1} \cdots B_{\alpha_q} = e^{-q} \sum_{k=1}^{\infty} \frac{q^k k^n}{k!} = \sum_{r=1}^{n} q^r S_n^r. \tag{5.5}$$

REFERENCES

L. COMTET (1974), *Advanced Combinatorics*, D. Reidel, Dordrecht, the Netherlands.

G. M. CONSTANTINE (1987), *Combinatorial Theory and Statistical Design*, John Wiley, New York.

C. S. CHEN AND T. H. SAVITS (1990), *Compound Nonhomogeneous Poisson Processes*, Tech. Report 90-03, Series in Reliability and Statistics, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA.

——— (1993), *Some remarks on compound nonhomogeneous Poisson processes*, Statist. Probab. Lett., 17, pp. 179–187.

G. DOBINSKI (1877), Grunert's Archiv., 61, pp. 333–336.

S. ROSS (1983), *Stochastic Processes*, John Wiley, New York.

G.-C. ROTA (1964), *The number of partitions of a set*, Ann. Math. Monthly, 71, pp. 498–503.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 203-212, May 1994

006

THE SQUARE OF A HAMILTONIAN CYCLE*

GENGHUA FAN[†] AND ROLAND HÄGGKVIST[‡]

Abstract. Let C be a cycle. The square of C is the graph obtained by joining every pair of vertices of distance 2 in C. Let G be a graph on n vertices with minimum degree $\delta(G)$. This paper proves that, if $\delta(G) \geq \frac{5}{7}n$, then G contains the square of a Hamiltonian cycle.

Key words. cycles, square of a cycle, Hamiltonian graphs

AMS subject classifications. 05C38, 05C45

1. Introduction. All graphs considered are simple and undirected. We denote by $\delta(G)$ the minimum degree of a graph G. A *k-chord* of a cycle C is an edge joining two vertices of distance k in C; the kth power of C is the graph obtained from C by joining every pair of vertices with distance at most k in C. As a natural generalization of Dirac's theorem [1], Seymour conjectured [7] that, if G is a graph on n vertices and $\delta(G) \geq (k/k + 1)n$, then G contains the kth power of a Hamiltonian cycle. This problem has been studied by Faudree et al. [4]. As mentioned by Seymour, the truth of his conjecture would imply the remarkably difficult Hajnal–Szemerédi theorem [6], which states that, if a graph G on n vertices has maximum degree less than r, then G is r-colorable such that the size of the color classes are all $\lfloor n/r \rfloor$ or $\lceil n/r \rceil$. It is clear that the case where $k = 1$ in the conjecture is Dirac's theorem. In this paper, we are only concerned with the case where $k = 2$. We use "square" instead of "2nd power." This special case was earlier conjectured by Pósa (see Erdös [2]).

CONJECTURE. *Let G be a graph on n vertices. If $\delta(G) \geq \frac{2}{3}n$, then G contains the square of a Hamiltonian cycle.*

The above conjecture, if true, is the best possible in the following sense. Let G be the complete 3-partite graph with one part containing m vertices and each of the other two parts containing $m + 1$ vertices. It is clear that $\delta(G) = 2m + 1 = \frac{2}{3}n - \frac{1}{3}$, but G does not contain the square of a Hamiltonian cycle. Jacobson (in a personal communication) showed that, if $\delta(G) \geq \frac{5}{6}n$, then the conclusion of the conjecture holds. By using a result in [5], Häggkvist (unpublished) gave a very simple proof of the cases where $\delta(G) \geq \frac{3}{4}n + 1$ and $n \equiv 0 \pmod 4$. In this paper, we prove the following result.

THEOREM. *Let G be a graph on n vertices. If $\delta(G) \geq \frac{5}{7}n$, then G contains the square of a Hamiltonian cycle.*

We need more notation and terminology. Define a k-chord of a path in a way similar to that of a cycle. A *feasible path* (*feasible cycle*) of a graph G is a path (cycle) containing all the 2-chords. A *strongly feasible* path is a feasible path, which, in addition, contains all the 3-chords. Denote by xy the edge with ends x and y. Two edges are *independent* if they have no common end. If $P = xyQzw$ is a feasible path in G, where Q is a path (so is feasible itself), then we say that xy is *feasibly connected* to zw by P, or that xy is *feasibly connected* to zw through Q, and we write $xy \twoheadrightarrow zw$ in G. In the case where $Q = \emptyset$, we simply say that xy is *feasibly joined* to zw. By reversing the order of the path, we see that $xy \twoheadrightarrow zw$ is equivalent to $wz \twoheadrightarrow yx$. However, $xy \twoheadrightarrow zw$ is not the same as $yx \twoheadrightarrow zw$. Here the order of the two ends of the end-edges of a feasible path is significant. This is

* Received by the editors June 10, 1992; accepted for publication (in revised form) March 23, 1993.

[†] Department of Mathematics, University of Umeå, Umeå, Sweden. Present address, Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

[‡] Department of Mathematics, University of Umeå, Umeå, Sweden.

the only case where we distinguish xy from yx for the same edge with ends x and y. In such a case, to avoid vagueness, we use *ordered edges* instead of edges. Let $C = u_1u_2\cdots u_mu_1$ be a cycle. We say that u_i and u_{i+1} are consecutive on C, $1 \le i \le m$, where the additions of the indices of u_i are taken modulo m. We choose an orientation of C that is consistent with the increasing order of the indices of u_i $(1 \le i \le m)$. By $u_i < u_j$, we mean that $u_i \ne u_j$, and there is a path of C from u_i to u_j, determined by the orientation of C. We call this path the *segment* of C from u_i to u_j. We use K_m for the complete graph on m vertices. The vertex set of a graph G is denoted by $V(G)$ and the edge set by $E(G)$. If H is a subgraph of G, we denote by $G - H$ the subgraph obtained by deleting all the vertices of H together with all the edges, with at least one end in H; for $v \in V(G)$, $N_H(v)$ is the set, and $d(v, H)$ the number, of vertices in H that are joined to v. We call $d(v, H)$ the degree of v in H and we define $d(F, H) = \sum_{v\in V(F)} d(v, H)$, for two subgraphs F and H of G. By definition, $d(F, H) = d(H, F)$, and, if $V(F) \cap V(H) = \varnothing$, then $d(F, H)$ is the number of edges with one end in F and other end in H.

Results on feasible paths and cycles are proved in §§2 and 3, respectively. They are applied to the proof of the main theorem in the last section. The proof of the theorem is by induction on the number of vertices of the graphs. We choose a longest feasible cycle C with maximum number of 3-chords. Set $H = G - C$ and define $c = |V(C)|$. We first show that $d(x, C) \le \frac{5}{7}c$ for every vertex x in H, and so, by induction, $c \ge n/2$. Next, we prove that $d(T, C) \le 2c$ for every triangle T in H, which enables us to establish that every two independent ordered edges of H are feasibly connected in H. Moreover, we prove that H contains at least $\frac{3}{7}n$ vertices. By these results, we find that the subgraph $G - H$ (induced by $V(C)$) has such a high density that every two independent edges of C are connected by a feasible path of at least $c - 2$ vertices in $G - H$, which, together with a feasible path in H, results in a feasible cycle of vertices more than c. This gives a contradiction to the choice of C.

2. Results on feasible paths.

LEMMA 2.1. *Let $P = x_1x_2\cdots x_p$ be a strongly feasible path in a graph G, where $p \ge 5$. Denote by T the triangle induced by $\{x_{p-2}, x_{p-1}, x_p\}$. Then x_1x_2 is connected to each of the six ordered edges of T by a feasible path P' with $V(P') \subseteq V(P)$ and $|V(P')| \ge p - 1$.*

Proof. Evidently, the path P itself gives the required paths from x_1x_2 to the ordered edges $x_{p-2}x_{p-1}$ and $x_{p-1}x_p$. Since P is strongly feasible and $p \ge 5$, we have that $x_{p-3}x_p \in E(G)$, and so the path $x_1x_2\cdots x_{p-3}x_{p-2}x_px_{p-1}$ is feasible, which provides the required paths from x_1x_2 to $x_{p-2}x_p$ and x_px_{p-1}. Moreover, since $x_{p-4}x_{p-1} \in E(G)$, both $x_1x_2\cdots x_{p-3}x_{p-1}x_{p-2}$ and $x_1x_2\cdots x_{p-3}x_{p-1}x_px_{p-2}$ are feasible paths of G. This completes the proof. □

LEMMA 2.2. *Let $P = x_1x_2\cdots x_p$ be a path in a graph G and $ab \in E(G - P)$. If neither ab nor ba is feasibly joined to any ordered edge $x_ix_{i+1} \in E(P)$, $1 \le i \le p - 1$, then $d(ab, P) \le p + 1$, and, furthermore, if $d(x_p, ab) \le 1$, then $d(ab, P) \le p$ with equality only if $d(x_1, ab) \ge 1$.*

Proof. Set $N = \{x \in V(P) : d(x, ab) = 2\}$ and $N^+ = \{x_{i+1} \in V(P) : x_i \in N\}$. By the given condition, if $x_i \in N$, then $d(x_{i+1}, ab) = 0$. Thus $N \cap N^+ = \varnothing$ and $d(ab, N \cup N^+) \le 2|N|$. Noting that $|N \cup N^+| \ge 2|N| - 1$ and since $d(x, ab) \le 1$ for all $x \in V(P)\backslash(N \cup N^+)$, we have that $d(ab, P) \le 2|N| + (p - |N \cup N^+|) \le p + 1$, as required. Furthermore, if $d(x_p, ab) \le 1$, then $|N \cup N^+| = 2|N|$, and so $d(ab, P) \le p$, with equality only if $d(x, ab) = 1$ for all $x \in V(P)\backslash(N \cup N^+)$, which implies, since $x_1 \notin N^+$, that $d(x_1, ab) \ge 1$. □

LEMMA 2.3. *Let $P = x_1x_2\cdots x_p$ be a path in a graph G. Suppose that T is a triangle in $G - P$. If none of the six ordered edges of T is feasibly joined to any ordered edge*

$x_ix_{i+1} \in E(P)$, $1 \le i \le p-1$, *then* $d(T, P) \le \frac{3}{2}(p+1)$, *and, furthermore, if* $d(x_p, T) \le 1$, *then* $d(T, P) \le \frac{3}{2}p$ *with equality only if* $d(x_1, T) \ge 2$.

Proof. Apply Lemma 2.2 to each edge of T. □

LEMMA 2.4. *Suppose that* G *is a graph on* n *vertices in which* $d(T, G) > \frac{9}{4}n$ *for every triangle* T *of* G. *If* f *is an edge of* G *contained in a triangle, then it is contained in a* K_5.

Proof. Let T be a triangle containing f. Since $d(T, G) > \frac{9}{4}n$, there is $v \in G - T$ such that $T \cup \{v\}$ induces a K_4. Let Q denote such a K_4. Applying the given condition to every triangle of Q, we have that $d(Q, G) > 3n$. If $d(v, Q) \le 3$ for all $v \in V(H - Q)$, then $d(Q, G) \le 3(n-4) + 12 = 3n$, a contradiction. Therefore, $d(v, Q) = 4$ for some $v \in V(G - Q)$, which gives a K_5 containing f.

LEMMA 2.5. *Let* K *be a complete subgraph of* G *and* xy *an ordered edge of* K. *If* $ab \in E(G - K)$, *with that order, and* $d(ab, K) \ge |V(K)| + 3$, *then* $xy \twoheadrightarrow ab$ *in* G.

Proof. Since $d(ab, K) \ge |V(K)| + 3$, there are $u, v \in V(K)\backslash\{x\}$ such that $d(uv, ab) = 4$. So $xy \twoheadrightarrow uv \twoheadrightarrow ab$, where we choose $u = y$ if $y \in \{u, v\}$.

LEMMA 2.6. *Suppose that* G *is a graph on* n *vertices in which* $d(T, G) > \frac{9}{4}n$ *for every triangle* T *of* G. *Let* xy *and* zw *be any two independent ordered edges. If both* xy *and* zw *are contained in triangles, then* $xy \twoheadrightarrow zw$ *in* G.

Proof. Suppose, to the contrary, that this is not true. By Lemma 2.4, xy is contained in some K_5 in G. This implies that xy is contained in some triangle in $G - zw$. Let K be a maximal complete subgraph containing xy in $G - zw$. Set $k = |V(K)|$. By the maximality of K, $d(K, G - K - zw) \le (k-1)(n-k-2)$, and, by Lemma 2.5, $d(K, zw) \le k + 2$. Therefore, $d(K, G) \le (k-1)n - k + 4$. On the other hand, applying the condition that $d(T, G) > \frac{9}{4}n$ to every triangle of K, we have that $d(K, G) > (3k/4)n$. So $k \ge 5$. Similarly, there is also a complete subgraph on at least five vertices in $G - xy$, which contains zw. Therefore, we may let K_f and K_g be two K_5 containing xy and zw, respectively, such that $V(K_f) \cap V(K_g) \subseteq V(G)\backslash\{x, y, z, w\}$. If $|V(K_f) \cap V(K_g)| \ge 2$, then xy and zw are connected by a feasible path through an edge in $K_f \cap K_g$, contradicting the assumption. Suppose therefore that $|V(K_f) \cap V(K_g)| \le 1$.

If $|V(K_f) \cap V(K_g)| = 1$, say $V(K_f) \cap V(K_g) = \{u\}$, we denote by $Q_f(Q_g)$ the K_4 obtained by deleting u from $K_f(K_g)$. If there is an edge e from $V(Q_f)\backslash\{x\}$ to $V(Q_g)\backslash\{w\}$, then xy and zw are connected by a feasible path through the vertex u with e as a 2-chord, a contradiction again. This shows that $d(Q_f, Q_g) \le 7$. Let $R = G - K_g - K_f$. If $d(v, Q_f) \le 3$ for all $v \in V(R)$, then $d(Q_f, R) \le 3(n-9)$, and so

$$\begin{aligned} d(Q_f, G) &= d(Q_f, R) + d(Q_f, Q_g) + d(Q_f, u) + d(Q_f, Q_f) \le 3(n-9) + 7 + 4 + 12 \\ &= 3n - 4. \end{aligned}$$

However, by applying the given condition to every triangle of Q_f, we have that $d(Q_f, G) \ge 3n$. This contradiction shows that there must be a vertex $v \in V(R)$ such that $d(v, Q_f) = 4$, and then Q_f together with v induces a K_5 vertex-disjoint with K_g.

From the above discussion, we see that either there is a K_5 that contains xy and is vertex-disjoint with K_g, or $V(K_f) \cap V(K_g) = \emptyset$. Therefore, we may choose K_f, in the first place, such that $V(K_f) \cap V(K_g) = \emptyset$. Now let

$$P_f = xya_1a_2\cdots a_{p-2}a_{p-1}a_p \quad \text{and} \quad P_g = zwb_1b_2\cdots b_{q-2}b_{q-1}b_q$$

be two vertex-disjoint strongly-feasible paths starting at xy and zw such that (i) $p \ge 3$ and $q \ge 3$, and (ii) subject to (i), $p + q$ is maximum. Such two paths exist from the existence of K_f and K_g. Let T_f and T_g be the triangles induced by the last three vertices of P_f and P_g, respectively. Set $R = G - P_f - P_g$. By the maximality of $p + q$, $d(T_f, v) \le 2$ for all $v \in V(R)$, which implies that $d(T_f, R) \le 2(n - |V(P_f)| - |V(P_g)|)$. By Lemma 2.1, if there is an ordered edge of T_f (T_g) that is feasibly joined to an ordered edge

$b_ib_{i-1} \in E(P_g - zw)$ $(a_ia_{i-1} \in E(P_f - xy))$, then $xy \twoheadrightarrow zw$, and we are done. (Note that zb_2, $xa_2 \in E(G)$.) Suppose that this is not the case. Applying Lemma 2.3 to T_f and $P_g - zw$ (with reversed order on P_g), we have that $d(T_f, P_g - zw) \leq \frac{3}{2}(q+1)$. Clearly, if $d(T_f, zw) \geq 5$, then $xy \twoheadrightarrow zw$. Suppose that $d(T_f, zw) \leq 4$. Consequently,

$$d(T_f, G) \leq 2(n - |V(P_f)| - |V(P_g)|) + 3(|V(P_f)| - 1) + \tfrac{3}{2}(q+1) + 4.$$

Using $q = |V(P_g)| - 2$,

$$d(T_f, G) \leq 2n + |V(P_f)| - \tfrac{1}{2}|V(P_g)| - \tfrac{1}{2}.$$

By symmetry,

$$d(T_g, G) \leq 2n + |V(P_g)| - \tfrac{1}{2}|V(P_f)| - \tfrac{1}{2}.$$

Adding the last two inequalities together and using $|V(P_f)| + |V(P_g)| \leq n$, we find that $d(T_f, G) + d(T_g, G) \leq \frac{9}{2}n - 1$. However, by the given condition, $d(T_f, G) + d(T_g, G) > \frac{9}{2}n$. This contradiction proves the lemma. □

3. Results on feasible cycles. Throughout this section, G denotes a graph with minimum degree $\delta(G) \geq \frac{5}{7}n$, where n is the number of vertices of G, and C is a longest feasible cycle of G: $C = u_1u_2u_3\cdots u_{c-1}u_cu_1$, where $c = |V(C)|$. Set $H = G - C$ and $h = |V(H)|$, where $h > 0$. Note that $\delta(G) \geq \frac{5}{7}n$ implies $c \geq 4$.

LEMMA 3.1. *If $v \in V(H)$, then no four consecutive vertices of C are all joined to v.*

Proof. Consider any four consecutive vertices $\{u_i, u_{i+1}, u_{i+2}, u_{i+3}\}$ of C. If v is joined to all of them, then $u_iu_{i+1}vu_{i+2}u_{i+3}$ is a feasible path, which implies that we can insert v into C and obtain a longer feasible cycle, contradicting the choice of C. □

LEMMA 3.2. *Let T be a triangle of H with $V(T) = \{x, y, z\}$. If there are independent edges u_iu_{i+1}, $u_ju_{j+1} \in E(C)$, $j > i$, such that $d(u_iu_{i+1}, xy) = 4$ and $d(u_ju_{j+1}, z) = 2$, then either $d(u_j, xy) = 0$ or $j \geq i + 5$.*

Proof. If u_j is joined to $x(y)$, then $u_iu_{i+1}yxzu_ju_{j+1}$ $(u_iu_{i+1}xyzu_ju_{j+1})$ is a feasible path, which yields a feasible cycle of length $c + 3 - (j - i - 2) = c + i + 5 - j$. By the maximality of C, $j \geq i + 5$. □

LEMMA 3.3. *Let T be a triangle of H. For any i, $1 \leq i \leq c$, if $d(u_iu_{i+1}, T) \geq 5$, then $d(u_{i+2}u_{i+3}, T) \leq 4$ with equality only if $d(u_{i+2}, T) = 1$.*

Proof. If $d(u_{i+2}u_{i+3}, T) \geq 5$, then $d(\{u_i, u_{i+1}, u_{i+2}, u_{i+3}\}, T) \geq 10$, which implies that some vertex of T is joined to all of the four consecutive vertices u_{i+l}, $0 \leq l \leq 3$. This is impossible by Lemma 3.1. Thus $d(u_{i+2}u_{i+3}, T) \leq 4$. If the equality holds, then there is a vertex $z \in V(T)$ such that $d(z, u_{i+2}u_{i+3}) = 2$. Furthermore, since $d(u_iu_{i+1}, T) \geq 5$, there is an edge $xy \in E(T)$ such that $d(xy, u_iu_{i+1}) = 4$. By Lemma 3.1, $z \notin \{x, y\}$. It follows from Lemma 3.2 that $d(u_{i+2}, xy) = 0$, and hence $d(u_{i+2}, T) = 1$. □

Reversing the order of the indices in Lemma 3.3, we have the following result.

LEMMA 3.4. *Let T be a triangle of H. For any i, $1 \leq i \leq c$, if $d(u_iu_{i-1}, T) \geq 5$, then $d(u_{i-2}u_{i-3}, T) \leq 4$ with equality only if $d(u_{i-2}, T) = 1$.*

LEMMA 3.5. *Let T be a triangle of H. Suppose that $a_1b_1a_2b_2\cdots a_kb_k$ is a segment of C such that $d(a_ib_i, T) = 4$ for all i, $1 \leq i \leq k$. If $d(a_1, T) = 1$, then $d(b_k, T) = 3$.*

Proof. If $k = 1$, the assertion is trivially true. We proceed by induction on k. Since $d(a_1b_1, T) = d(a_2b_2, T) = 4$, there are $x, y \in V(T)$ such that x is joined to both a_1 and b_1 and y to both a_2 and b_2. By Lemma 3.1, $x \neq y$. Since $d(a_1, T) = 1$ and $d(a_1b_1, T) = 4$, we have that $d(b_1, T) = 3$. If $a_2x \in E(G)$, then $a_1b_1xya_2b_2$ is a feasible path; if $a_2z \in E(G)$, where z is the third vertex of T other than x and y, then $a_1b_1xzya_2b_2$ is a feasible path. Both cases lead to a feasible cycle of length at least $c + 2$. This contradiction shows that y is the only vertex of T joined to a_2. That is, $d(a_2, T) = 1$. Applying the induction hypothesis to $a_2b_2a_3b_3\cdots a_kb_k$ completes the proof of the lemma. □

LEMMA 3.6. *Let T be a triangle of H. Then $d(T, C) \leq 2c$.*

Proof. If $c = 4$, by the maximality of c, $d(T, C) \leq 7$. If $c = 5$, by Lemma 3.1, $d(T, C) \leq 9$. Suppose that $c \geq 6$. Let M be a perfect matching of $E(C)$ if c is even or a perfect matching of $C - u$ otherwise, where $u \in V(C)$ chosen so that $d(u, T) = \min\{d(v, T) : v \in V(C)\}$. Set $M = \{e_1, e_2, \ldots, e_m\}$, where $m = \lfloor c/2 \rfloor$ and e_i, e_{i+1} are consecutive edges on C, $1 \leq i \leq m - 1$. If $d(e_i, T) \leq 4$ for all i, $1 \leq i \leq m$, then $d(M, T) \leq 4m$, which gives that $d(C, T) = 2c$, and we are done. Therefore, suppose that the set

$$D = \{e_i \in M : d(e_i, T) \geq 5, 1 \leq i \leq m\}$$

is not empty and divides C into segments. A typical member of these segments is of form

$$S = a_0b_0a_1b_1 \cdots a_kb_k \quad \text{or} \quad S = a_0b_0a_1b_1 \cdots a_qb_qua_{q+1}b_{q+1} \cdots a_kb_k,$$

depending on whether $u \in V(S)$, where $a_ib_i \in M$, $0 \leq i \leq k$ with $a_0b_0 \in D$, but $a_jb_j \notin D$ for all j, $1 \leq j \leq k$; subject to these, k is as large as possible. If $|D| \geq 2$, let $f \in M - E(S)$ be the edge next to a_kb_k on C, and, by the maximality of k, $d(f, T) \geq 5$. Noting that $d(a_0b_0, T) \geq 5$ and applying Lemma 3.3, we have that $k \geq 1$. If $|D| = 1$, let $f = a_0b_0$. Since $c \geq 6$, we have that $k \geq 2$. By the construction of S,

$$(3.1) \qquad d(a_jb_j, T) \leq 4 \quad \text{for all } j, \quad 1 \leq j \leq k.$$

We prove that

$$(3.2) \qquad d(S, T) \leq 2|V(S)|$$

by using

$$(3.3) \qquad d(S, T) = \sum_{i=0}^{k} d(a_ib_i, T) \quad \text{or} \quad d(S, T) = \sum_{i=0}^{k} d(a_ib_i, T) + d(u, T).$$

Case (i). $u \notin V(S)$. We first consider the subcase in which $d(a_0b_0, T) = 5$. If all equalities hold in (3.1), then $d(a_1, T) = 1$ by Lemma 3.3, and so $d(b_k, T) = 3$ by Lemma 3.5. However, by Lemma 3.4 with $d(f, T) \geq 5$, $d(b_k, T) = 1$. This contradiction shows that there must be a strict inequality in (3.1), and then (3.2) follows from (3.3). Next, consider the subcase in which $d(a_0b_0, T) = 6$. Since $d(f, T) \geq 5$ and by Lemma 3.2, we have that $k \geq 2$. Since $d(a_0b_0, T) = 6$ and by Lemma 3.1, no vertex of T can be joined to both a_1 and b_1, and hence $d(a_1b_1, T) \leq 3$. Now, if there is some t, $2 \leq t \leq k$, such that $d(a_tb_t, T) \leq 3$, then (3.2) follows easily. Suppose that this is not the case. So $d(a_jb_j, T) = 4$ for all j, $2 \leq j \leq k$. Note that $d(a_2b_2, T) = 4$ implies that there is $z \in V(T)$ such that $d(a_2b_2, z) = 2$; then, by Lemma 3.2, $d(a_2, T - z) = 0$, and so $d(a_2, T) = 1$. Then, by Lemma 3.5, $d(b_k, T) = 3$. Again, however, by Lemma 3.4 with $d(f, T) \geq 5$, $d(b_k, T) = 1$. This contradiction proves (3.2). Summing (3.2) over all the segments S yields $d(C, T) \leq 2c$ and completes the proof of case (i).

Case (ii). $u \in V(S)$. Note that $d(a_0b_0, T) \leq 6$. If there is a strict inequality in (3.1), then $d(u, T) \leq 1$, and (3.2) follows from a simple counting. It remains to consider the subcase that all equalities hold in (3.1). By Lemmas 3.3 and 3.4, either $d(a_1, T) = 1$ (if $q \neq 0$) or $d(b_k, T) = 1$ (if $q \neq k$), either of which implies that $d(u, T) \leq 1$. If $d(a_0b_0, T) = 5$, then (3.2) follows. If $d(a_0b_0, T) = 6$, then, since $d(a_1b_1, T) = 4$, u must lie between b_0 and a_1, and moreover, by Lemma 3.2, $d(a_1, T) = 1$. Then, as in (i), we have the contradiction: $d(b_k, T) = 3$ and $d(b_k, T) = 1$. This proves (3.2) and the lemma. □

4. Proof of the theorem. Obviously, the assertion is true for complete graphs. We may assume that $n \geq 7$ and proceed by induction on n. Let C be a longest feasible cycle of G with maximum number of 3-chords: $C = u_1u_2u_3 \cdots u_{c-1}u_cu_1$, where $c = |V(C)|$.

(Note that C has the following additional property: maximum number of 3-chords.) As before, set $H = G - C$ and $h = |V(H)|$. If $h = 0$, there is nothing to prove. Assume that $h > 0$. Since G contains K_4, we have $c \geq 4$. If $c = 4$, then $d(x, C) \leq 3$ for every vertex in H, and the set $I = \{x \in V(H) : d(x, C) = 3\}$ forms an independent set. A simple counting yields a contradiction to the fact that $\delta(G) \geq \frac{5}{7}n$ and $n \geq 7$. In what follows, we assume that $c \geq 5$.

Claim 1. $d(x, C) \leq \frac{5}{7}c$ for every $x \in V(H)$.

If $c = 5$ or 6, the claim follows directly from Lemma 3.1. Suppose that $c \geq 7$. Let $x \in V(H)$ and $S = u_i u_{i+1} \cdots u_{i+6}$ be any segment of C with seven consecutive vertices. We show that $d(x, S) \leq 5$. If this is not true, by Lemma 3.1, the only possibility is that $u_{i+k}x \in E(G)$ for all k, $k \neq 3$ and $0 \leq k \leq 6$. Replacing u_{i+3} by x, we obtain a feasible cycle $C' = u_1 \cdots u_i u_{i+1} u_{i+2} x u_{i+4} u_{i+5} u_{i+6} \cdots u_c u_1$ of the same length as C. Note that $u_i x$ and $x u_{i+6}$ are 3-chords of C'. By the choice of C (containing maximum number of 3-chords), both $u_i u_{i+3}$ and $u_{i+3} u_{i+6}$ are in $E(G)$. So $u_1 \cdots u_i u_{i+1} u_{i+3} u_{i+2} u_{i+4} x u_{i+5} \times u_{i+6} \cdots u_c u_1$ is a feasible cycle of length $c + 1$, contradicting the choice of C. Summing $d(x, S) \leq 5$ over all the segments S completes the proof of Claim 1. An immediate consequence of Claim 1 is that

$$\delta(H) \geq \delta(G) - \tfrac{5}{7}c \geq \tfrac{5}{7}h, \tag{4.1}$$

which implies, by the induction hypothesis, that H contains the square of a Hamiltonian cycle of H. By the choice of C, $c \geq h$.

Claim 2. If T is a triangle of H, then $d(T, H) > \frac{9}{4}h$ in H.

By Lemma 3.6, $d(T, H) \geq 3\delta(G) - d(T, C) \geq \frac{15}{7}n - 2c = 2h + \frac{1}{7}n$. Since $c \geq h$, we have that $n \geq 2h$, and therefore $d(T, H) > \frac{9}{4}h$.

Claim 3. $h \geq \frac{3}{7}n$.

Let K be a maximum complete subgraph of H and set $k = |V(K)|$. By (4.1), $k \geq 4$, and, by the maximality of K, $d(K, v) \leq k - 1$ for all $v \in H - K$. Thus $d(K, H) \leq h(k - 1)$. Therefore

$$d(k, C) \geq k\delta(G) - d(k, H) \geq \tfrac{5}{7}kn - h(k - 1). \tag{4.2}$$

If $d(K, u_i u_{i+1}) \leq k + 1$ for every edge $u_i u_{i+1} \in E(C)$, then summing these inequalities over all i, $1 \leq i \leq c$ yields that $d(K, C) \leq ((k + 1)/2)c = ((k + 1)/2)(n - h)$. Combining this with (4.2) and using $k \geq 4$, we obtain that

$$h \geq \frac{3k - 7}{7(k - 3)} n > \frac{3}{7} n,$$

as required by the claim. Therefore assume that there is some edge $u_t u_{t+1} \in E(C)$ such that $d(K, u_t u_{t+1}) \geq k + 2$. By relabeling, we may assume that $t = 1$. That is, there is some edge $x_1 x_2 \in E(K)$ such that $d(x_1 x_2, u_1 u_2) = 4$. Let

$$P = x_1 x_2 \cdots x_{p-2} x_{p-1} x_p$$

be a strongly feasible path in H starting at $x_1 x_2$ with p maximum. By Claim 2 and Lemma 2.4, $p \geq 5$. Denote by T the triangle induced by $\{x_{p-2}, x_{p-1}, x_p\}$. Let

$$Q = u_3 u_4 \cdots u_{k+2} \quad \text{and} \quad Q' = u_c u_{c-1} \cdots u_{c-k'+1}$$

be the two segments starting at $u_3 u_4$ and $u_c u_{c-1}$ (Q' takes the reversed orientation of C), respectively, such that

(i) $V(Q) \cap V(Q') = \varnothing$ and both $k, k' \leq p$,
(ii) subject to (i), $d(u_{k+2}, T) \leq 1$ and $d(u_{c-k'+1}, T) \leq 1$,
(iii) subject to (i) and (ii), $k + k'$ is maximum.

Note that $|V(Q)| = k \leq p$ and $|V(Q')| = k' \leq p$. Applying Lemma 2.1 to P and by the maximality of C, no ordered edge of T is feasibly joined to $u_i u_{i+1} \in E(Q)$ or $u_j u_{j-1} \in E(Q')$. By choice (ii) of Q and Q', it follows from Lemma 2.3 that

$$d(T, Q) + d(T, Q') \leq \tfrac{3}{2}(k + k') \tag{4.3}$$

with equality only if both $d(u_3, T) \geq 2$ and $d(u_c, T) \geq 2$, and from Lemma 2.2 that

$$d(x_{p-1}x_p, Q) + d(x_{p-1}x_p, Q') \leq k + k' \tag{4.4}$$

with equality only if both $d(u_3, x_{p-1}x_p) \geq 1$ and $d(u_c, x_{p-1}x_p) \geq 1$. Since no ordered edge of T is feasibly joined to $u_i u_{i+1}$ ($u_j u_{j-1}$), we have that either $d(u_i, T) \leq 1$ or $d(u_{i+1}, T) \leq 1$ (either $d(u_j, T) \leq 1$ or $d(u_{j-1}, T) \leq 1$), for any i (j) with $3 \leq i \leq p + 1$ ($c - k' + 2 \leq j \leq c$). This implies that we may choose Q and Q' such that either (i) $k + k' \geq 2p - 2$ if $c \geq 2p + 1$ or (ii) $c - 3 \leq k + k' \leq c - 2$ if $c \leq 2p$. We discuss these two cases separately. For convenience, set $R = C - Q - Q' - u_1 u_2$.

Case (i). $c \geq 2p + 1$ and $k + k' \geq 2p - 2$. Note that $d(T, u_1 u_2) \leq 6$. By Lemma 3.1, if equality holds in (4.3), then $d(T, u_1 u_2) \leq 5$. Therefore

$$d(T, Q) + d(T, Q') + d(T, u_1 u_2) \leq \tfrac{3}{2}(k + k') + \tfrac{11}{2}. \tag{4.5}$$

If $|V(R)| \leq 2$, then clearly $d(T, R) \leq 2|V(R)| + 2 = 2c - 2(k + k') - 2$. Adding this to (4.5) and then using $k + k' \geq 2p - 2$, we obtain that

$$d(T, C) \leq 2c - p + \tfrac{9}{2}. \tag{4.6}$$

Furthermore, by the choice of P, $d(T, v) \leq 2$ for all $v \in V(H - P)$, and therefore $d(T, H) \leq 2(h - p) + 3(p - 1) = 2h + p - 3$. Consequently,

$$d(T, G) = d(T, C) + d(T, H) \leq 2n + \tfrac{3}{2}. \tag{4.7}$$

On the other hand, $d(T, G) \geq 3\delta(G) \geq \frac{15}{7}n$. It follows that $n \leq 10$. However, $c \geq 2p + 1$ implies that $n \geq 3p + 1 \geq 16$, a contradiction. Suppose now that $|V(R)| \geq 3$. From the proof of Lemma 3.6 (see (3.2)), it is not difficult to see that $d(T, R) \leq 2|V(R)| + 3 = 2c - 2(k + k') - 1$. Adding this to (4.5), instead of (4.6) we have that $d(T, C) \leq 2c - p + \frac{11}{2}$, and, by the same argument used in deriving (4.7), $d(T, G) \leq 2n + \frac{5}{2}$, which, together with $d(T, G) \geq \frac{15}{7}n$, implies that $n \leq 17$. However, using $k + k' \geq 2p - 2$ and $|V(R)| \geq 3$, we have that $c = k + k' + 2 + |V(R)| \geq 2p + 3$, and so $n \geq 3p + 3 \geq 18$, a contradiction again.

Case (ii). $c \leq 2p$ and $c - 3 \leq k + k' \leq c - 2$. Then $|V(R)| \leq 1$. Consider the edge $x_{p-1}x_p$. Note that $d(x_{p-1}x_p, u_1 u_2) \leq 4$. If equality holds in (4.4), then $d(x_{p-1}x_p, u_1 u_2) \leq 3$, for otherwise, either Lemma 3.1 is violated or the edge $x_{p-1}x_p$ (or $x_p x_{p-1}$) can be inserted into C between u_1 and u_2 to obtain a feasible cycle of length $c + 2$. It follows that

$$d(x_{p-1}x_p, Q) + d(x_{p-1}x_p, Q') + d(x_{p-1}x_p, u_1 u_2) \leq k + k' + 3 = c - |V(R)| + 1.$$

Using $d(x_{p-1}x_p, R) \leq 2|V(R)|$, we derive that $d(x_{p-1}x_p, C) \leq c + |V(R)| + 1 \leq c + 2$. Therefore

$$d(x_{p-1}x_p, G) = d(x_{p-1}x_p, C) + d(x_{p-1}x_p, H) \leq c + 2 + 2(h - 1) = n + h.$$

On the other hand, $d(x_{p-1}x_p, G) \geq 2\delta(G) \geq \frac{10}{7}n$. It follows that $h \geq \frac{3}{7}n$, as required. This proves Claim 3.

Claim 4. $d(u, H) \geq \frac{1}{3}h + 1$ for every $u \in V(C)$.

We note that $d(u, H) \geq d(u) - (c - 1) \geq \frac{5}{7}n - c + 1 = h + 1 - \frac{2}{7}n$. From Claim 3, $n \leq \frac{7}{3}h$, and the claim follows.

Claim 5. Let ab and uv be any two ordered independent edges with both ends in C. If there are distinct vertices $x, y \in V(H)$ such that $d(x, ab) = 2$ and $d(y, uv) = 2$, then ab is feasibly connected to uv through a path P with $V(P) \subseteq V(H)$ and $|V(P)| \geq 3$.

If there are two distinct vertices $z, w \in V(H)\backslash\{x, y\}$ such that z is joined to both x, b and w to both y, u, then $ab \twoheadrightarrow xz$ and $wy \twoheadrightarrow uv$. By (4.1), xz is contained in a triangle in H, and so is wy. It follows from Claim 2 and Lemma 2.6 that $xz \twoheadrightarrow wy$, and so $ab \twoheadrightarrow uv$ through a path P with $V(P) \subseteq V(H)$ and $\{x, y, z, w\} \subseteq V(P)$. Suppose that this is not the case. From (4.1), $|N_H(x)| \geq \frac{5}{7}h > \frac{2}{3}h$. This, together with Claim 4, implies that b is joined to at least two vertices in $N_H(x)$. Similarly, u is joined to at least two vertices in $N_H(y)$. To avoid the case we just discussed, it must be that $xy \in E(G)$ and that there is $z \in N_H(x) \cap N_H(y)$ joined to both b and u such that $abxzyuv$ is a feasible path, which proves Claim 5 with $P = xzy$. We now return to the proof of our theorem.

Let M be a perfect matching of C if c is even or a perfect matching of $C - u$ otherwise, where $u \in V(C)$ chosen so that $d(u, H) = \min\{d(v, H) : v \in V(C)\}$. Set $M = \{e_1, e_2, \ldots, e_m\}$, where $m = \lfloor c/2 \rfloor$ and e_i, e_{i+1} are alternatively consecutive edges on C, $1 \leq i \leq m - 1$. (Note that, if c is odd, then u lies between e_m and e_1.)

First, consider the case where there are two distinct vertices $x, y \in V(H)$ such that $d(x, e_i) = 2$ and $d(y, e_j) = 2$ for two distinct edges $e_i, e_j \in M$. Since C is a longest feasible cycle and by Claim 5, we have $|i - j| \geq 2$, and moreover, $d(e_{i+1}, z) \leq 1$ for all $z \in V(H - x)$ and $d(e_{j+1}, z) \leq 1$ for all $z \in V(H - y)$. Furthermore, by Lemma 3.1, $d(e_{i+1}, x) \leq 1$ and $d(e_{j+1}, y) \leq 1$. Consequently,

$$(4.8) \qquad d(e_{i+1}, H) \leq h \quad \text{and} \quad d(e_{j+1}, H) \leq h.$$

Now we construct a new graph C^* with $V(C^*) = M$ as follows: Arrange $e_1, e_2, \ldots, e_m$ around a cycle in the same order as they are around C (so $e_t e_{t+1} \in E(C^*)$, $1 \leq t \leq m$) and add edges from e_{i+1} or e_{j+1} to e_l if and only if it is joined to e_l by four edges in G, $1 \leq l \leq m$. It is clear that any path of C^* gives rise to a feasible path of G with all vertices in C. Denote by α and β the degree of e_{i+1} and e_{j+1} in C^*, respectively. From the construction of C^*,

$$d(e_{i+1}, C) \leq 4\alpha + 3(m - 1 - \alpha) + d(e_{i+1}, C - M) + d(e_{i+1}, e_{i+1}).$$

Using $d(e_{i+1}, C - M) \leq 2(c - 2m)$ and $d(e_{i+1}, e_{i+1}) = 2$,

$$(4.9) \qquad d(e_{i+1}, C) \leq \alpha + 2c - m - 1.$$

On the other hand, using (4.8),

$$(4.10) \qquad d(e_{i+1}, C) \geq 2\delta(G) - d(e_{i+1}, H) \geq \tfrac{10}{7}n - h \geq c + \tfrac{3}{7}n.$$

By Claim 3, $c \leq \frac{4}{7}n$, that is, $n \geq \frac{4}{7}c$. Substituting this into (4.10) yields that $d(e_{i+1}, C) \geq \frac{7}{4}c$. This, together with (4.9), gives that $\alpha \geq m + 1 - \frac{1}{4}c$, but $c \leq 2m + 1$, and so

$$(4.11) \qquad \alpha \geq \frac{m}{2} + \frac{3}{4}.$$

Similarly, $\beta \geq m/2 + \frac{3}{4}$. Note that e_i and e_j are the predecessors of e_{i+1} and e_{j+1}, respectively. By a standard technique, e_i and e_j are connected by a Hamiltonian path in C^*, which implies that e_i and e_j are connected in G by a feasible path P with $V(P) \subseteq V(C)$ and $|V(P)| \geq c - 1$. However, from Claim 5, e_i and e_j (with any order of the ends) are connected through a feasible path P in H with $|V(P)| \geq 3$. The combination of these two paths yields a feasible cycle of G of length at least $c + 2$, a contradiction. Suppose therefore that this is not the case. It remains to consider the following two cases.

Case I. There is some vertex $x \in V(H)$ such that $d(x, e') = 2$ for some edge $e' \in M$. By relabeling, we may suppose that $d(x, e_1) = 2$. Then, to avoid the case we discussed above, it must be that no vertex in H other than x is joined to both ends of any edge of M other than e_1. Therefore

$$(4.12) \qquad d(e_i, H) \leq h + 1, \qquad 2 \leq i \leq m.$$

Now construct a graph M^* in which $V(M^*) = M$ and e_i is joined to e_j if and only if e_i and e_j are joined by four edges in G, $1 \leq i, j \leq m$. Denote by α_i the degree of e_i in M^*, $1 \leq i \leq m$. Using (4.12) instead of (4.8), we derive, instead of (4.11), that

$$(4.13) \qquad \alpha_i \geq \frac{m}{2} - \frac{1}{4}, \qquad 2 \leq i \leq m.$$

Since $d(x, e_1) = 2$, as shown in (4.8), we must have that $d(e_2, H) \leq h$, and therefore $\alpha_2 \geq m/2 + \frac{3}{4}$, which implies that $m \geq 4$ since $\alpha_2 \leq m - 1$. Note that the segment of C from e_m to e_2 through e_1 is a feasible path. If necessary, we may add two more edges to M^* from e_m to e_1 and from e_2 to e_1 so that $\alpha_1 \geq 2$. This, together with (4.13), implies that the graph M^* is 2-connected and max $\{d(u, M^*), d(v, M^*)\} \geq m/2$ for every pair of distinct vertices u and v in $V(M^*)$. By a result in [3], M^* has a Hamiltonian cycle, which implies a Hamiltonian path of M^* starting at e_1 with the last edge corresponding to a K_4 in G. This gives a feasible path P in G starting at e_1 such that $V(P) \subseteq V(C)$, $|V(P)| \geq c - 1$, and the last four vertices of P induce a K_4 in G, say $P = a_1b_1a_2b_2 \cdots a_{m-1}b_{m-1}a_mb_m$, where $a_1b_1 = e_1$. Since $a_{m-1}b_m \in E(G)$, the path P provides a feasible path containing at least $2m - 1$ vertices from a_1b_1 to each of the ordered edges $\{b_{m-1}a_m, b_{m-1}b_m, a_mb_m\}$. By Claim 4, there must be $y \in V(H - x)$ joined to both ends of one of these three ordered edges, and by Claim 5 such an ordered edge is feasibly connected to e_1 (with any order of the ends) through a path in H containing at least three vertices. Therefore, we have a feasible cycle of G of length at least $2m + 2 \geq c + 1$, a contradiction.

Case II. $d(x, e') \leq 1$ for every vertex $x \in V(H)$ and every edge $e' \in M$. Then $d(e_i, H) \leq h$ for all i, $1 \leq i \leq m$. Let M^* be the same graph as defined in Case I. Instead of (4.13), this time we have that

$$(4.14) \qquad \alpha_i \geq \frac{m}{2} + \frac{3}{4}, \qquad 1 \leq i \leq m.$$

So $m \geq 4$ and M^* is Hamiltonian. Let $P = v_1v_2 \cdots v_{p-1}v_p$ be the feasible path of G that corresponds to a Hamiltonian path of M^*. By the construction of M^*, $V(P) \subseteq V(C)$, $p = c$ or $c - 1$, and

$$(4.15) \qquad v_1v_4 \in E(G) \quad \text{and} \quad v_{p-3}v_p \in E(G).$$

If c is odd, then, by the choice of u, $d(u, H) \leq \frac{1}{2}h$, and so $d(u, P) \geq \frac{5}{7}n - \frac{1}{2}h = \frac{1}{2}c + \frac{3}{14}n$. Again, by Claim 3, $n \geq \frac{7}{4}c$, and thus $d(u, P) \geq \frac{7}{8}c$. Therefore, we may insert u into P in such a way that (4.15) still holds for the new path after the insertion. By this, we may assume that $p = c$ regardless of whether c is even or odd. Let $A = \{v_1v_2, v_1v_3, v_2v_3\}$ and $B = \{v_{p-2}v_{p-1}, v_{p-2}v_p, v_{p-1}v_p\}$ be two sets of ordered edges (with those given orders). Since $m \geq 4$, we have $V(A) \cap V(B) = \varnothing$. The path P with property (4.15) implies that every member of A is connected to every member of B by a feasible path P' with $V(P') \subseteq V(P)$ and $|V(P')| \geq |V(P)| - 2 = c - 2$. By Claim 4, there are distinct vertices $x, y \in V(H)$ such that x is joined to both ends of an ordered edge of A and y to both ends of an ordered edge of B. As seen in the proof of Case I, these two ordered edges are feasibly connected through a path in H containing at least three vertices, which, together with P', give a feasible cycle of G of length at least $c + 1$. This final contradiction completes our proof. □

REFERENCES

[1] G. A. DIRAC, *Some theorems on abstract graphs*, Proc. London Math. Soc., 2 (1952), pp. 68–81.

[2] P. ERDÖS, *Problem* 9, in Theory of Graphs and Its Applications, M. Fiedler, ed., Czech. Acad. Sci. Publ., Prague, 1964, p. 159.

[3] G. FAN, *New sufficient conditions for cycles in graphs*, J. Combin. Theory Ser. B, 37 (1984), pp. 221–227.

[4] R. J. FAUDREE, R. J. GOULD, M. S. JACOBSON, AND R. H. SCHELP, *On a problem of Paul Seymour*, in Recent Advances in Graph Theory, V. R. Kulli, ed., Vishwa, Oakland, CA, 1991, pp. 197–215.

[5] R. HÄGGKVIST, *On F-Hamiltonian graphs*, in Graph Theory and Related Topics, J. A. Bondy and U. S. R. Murty, eds., Academic Press, New York, 1979, pp. 219–231.

[6] A. HAJNAL AND E. SZEMERÉDI, *Proof of a conjecture of P. Erdös*, in Combinatorial Theory and Its Application, P. Erdös, A. Rênyi, and V. T. Sós, eds., North-Holland, London, 1970, pp. 601–623.

[7] P. SEYMOUR, *Problem section*, in Combinatorics: Proceedings of the British Combinatorial Conference 1973, T. P. McDonough and V. C. Mavron, eds., Cambridge University Press, Cambridge, UK, 1974, pp. 201–202.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 213-220, May 1994

007

AN $O(n \log n)$ ALGORITHM FOR BANDWIDTH OF INTERVAL GRAPHS*

ALAN P. SPRAGUE†

Abstract. This paper presents an $O(n \log n)$ algorithm for the bandwidth problem on interval graphs. Given an interval model for an interval graph with n vertices and an integer k, the algorithm constructs a layout of bandwidth at most k, if there exists one. Two previous algorithms for this problem have been published. One of them is flawed; the other is by Kleitman and Vohra and has complexity $O(nk)$.

Key words. interval graphs, bandwidth, graph algorithms

AMS subject classifications. 05C78, 05C85

1. Introduction. An *interval graph* is a graph whose vertices can be put in one-to-one correspondence with a family of intervals on the real line, such that two vertices are adjacent if and only if the corresponding intervals are nondisjoint. Such a family of intervals is called an *interval model* for the graph. Interval graphs have many applications [5], and algorithmic issues involving interval graphs have attracted much interest [3], [6], [7], [10].

Vertices v and w in a graph are said to be *neighbors* if $v = w$ or v is adjacent to w. A *layout* of a graph G on n vertices is a bijection between the vertices of G and the set $\{1, 2, \ldots, n\}$. In particular, we treat $\{1, 2, \ldots, n\}$ as the domain of the function and the vertices of G as the range. The *bandwidth* of layout λ equals max $\{|i - j| : \lambda(i)$ is a neighbor of $\lambda(j)\}$. The bandwidth of G is the minimum bandwidth of layouts of G.

Given a graph G and an integer k, the problem of determining if the bandwidth of G is at most k is called the *bandwidth problem.* The bandwidth problem is NP-complete on graphs and even on trees [11]. It is polynomial on caterpillars with hairlength at most 2 [1].

Kratsch [9] devised the first algorithm for bandwidth on interval graphs, but it is flawed; a counterexample is provided below. Kleitman and Vohra developed a different algorithm for this problem and proved it correct [8]. The algorithm of [8] has complexity $O(kn)$, where n is the number of vertices of the graph and k is the bandwidth being tested.

In this paper, we solve the bandwidth problem in $O(n \log n)$ time. The algorithm is essentially that of Kleitman and Vohra; appropriate data structures enable this $O(n \log n)$ running time, which is an improvement over their $O(kn)$ time for all interval graphs, except those with very low bandwidth.

We comment on the effect that the form of the input has on the complexity of our algorithm and other algorithms on interval graphs. In this paper, we assume that an interval model is given; under this assumption, our algorithm has complexity $O(n \log n)$. In some papers, it is assumed that an interval model is given and, in addition, that endpoints are already sorted. Under this assumption, we may aspire to $O(n)$ algorithms, which have been found for various problems, including maximum clique, clique cover, and maximum independent set [6], while our algorithm retains its $O(n \log n)$ complexity. On other occasions, it is assumed that an interval graph is represented as a consecutive cliques arrangement [9]. If we interpret this term to mean that, for each vertex, the first and last clique containing it is given, then, from a consecutive cliques arrangement, we

* Received by the editors June 15, 1992; accepted for publication (in revised form) March 23, 1993.

† Department of Computer and Information Sciences, University of Alabama, Birmingham, Alabama 35294.

may easily obtain an interval model in which endpoints of intervals are integers between 1 and n inclusive, which may be sorted in $O(n)$ time. Hence, given a consecutive cliques arrangement, $O(n)$ algorithms are known for various problems on interval graphs. Finally, it is sometimes assumed that an interval graph is represented in the usual adjacency list representation for general graphs. Such a representation has size $O(m + n)$, where m is the number of edges. The algorithm of Booth and Lueker [4] constructs an interval model for such a graph in $O(m + n)$ time, so our algorithm has complexity $O(m + n \log n)$, given an adjacency list representation for the graph.

In §2 we give an algorithm for bandwidth of interval graphs. In §§3 and 4 we develop data structures that enable the algorithm to be executed in $O(n \log n)$ time. The conclusion states several open problems.

2. Bandwidth. We begin this section by introducing some notation and display an example for which the algorithm of [9] errs. The bandwidth algorithm that we use is then stated. Finally, we explain the role that data structures must play for the complexity of the algorithm to be $O(n \log n)$.

We assume that an interval model for the interval graph G is given; for each interval z, z_L is the coordinate of the left end of z, and z_R its right end. There are n intervals. The arrays L_array and R_array may be filled in $O(n \log n)$ time with the left (respectively, right) ends of the intervals, in ascending order. (Ties may be broken arbitrarily.) We often speak of L_array and R_array as sequences of intervals rather than as sequences of coordinates. Array LR_array may be filled with the $2n$ ends in ascending order, by merging the other two. (Where intervals are closed intervals, whenever a left end and a right end have the same coordinate, the left end must precede the right end in LR_array.)

Kratsch [9] developed an approach that is used in [8] and this paper. A layout is constructed from left to right. This approach follows a policy of laying out intervals generally in order of right ends, but being mindful of constraints imposed by already-laid-down intervals. However, the first algorithm [9] to use this approach is incorrect. In particular, it erroneously decides that the interval graph of Fig. 1 has no layout of bandwidth 4. Property 2 of [9, p. 146] is a cause of the failure of this example.

The algorithm presented here may be viewed as the algorithm of Kleitman and Vohra, supported by new data structures. The following definitions are important in describing our concepts, data structures, and algorithm. Let Layout$(1 \cdots n)$ be the layout being constructed and suppose that intervals have been assigned to positions $1, 2, \ldots, t$. The set of intervals assigned to positions $1, 2, \ldots, t$ is written as Layout$(1 \cdots t)$. For any interval z that is not yet laid down, Deadline$(z) = \min(\{n\} \cup \{j + k : \text{Layout}(j)$ is adjacent to z, $1 \leq j \leq t\})$. Thus the deadline of z reflects the constraints on z imposed by the already-laid-down intervals. For $1 \leq h \leq n$, let I_h be the set of h intervals with leftmost left ends. Thus I_h corresponds to L_array$(1 \cdots h)$. Deadline$(I_h) = \max\{\text{Deadline}(z) : z \in I_h - \text{Layout}(1 \cdots t)\}$. We define Leeway$(I_h) = \text{Deadline}(I_h) - h - |\text{Layout}(1 \cdots t) - I_h|$; to understand this, note that $|\text{Layout}(1 \cdots t) - I_h|$ is the number

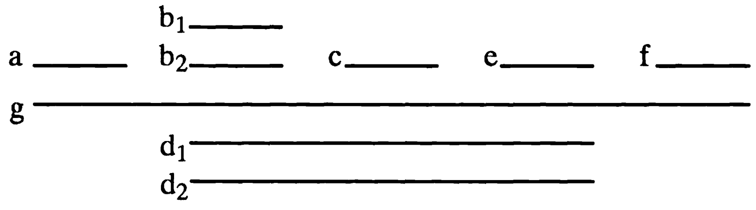


FIG. 1. *Model of interval graph, providing a counterexample to the algorithm of* [9].

of intervals outside I_h that are already laid down and that Deadline(I_h) $-$ h is the number of intervals outside I_h that will have been laid down by the time the deadline for I_h occurs; so Leeway(I_h) is the number of intervals outside I_h that are allowed to be laid down between now and the time of laying down the last interval of I_h. However, if all intervals of I_h are already laid down, we define Leeway(I_h) to be infinite. If Leeway(I_h) $= 0$, we say that I_h is *critical.* Because each not-yet-laid-down interval is defined to have a deadline that is no greater than n, there is always at least one critical set, for I_n is critical.

Observe that, if I_h is critical, the next interval to be laid down must be from I_h, to construct a layout of the desired bandwidth. Since the sets I_h are nested ($I_h \subset I_{h+1}$ for all h), choosing the next interval to be laid down from the smallest critical set assures that progress will be made on laying down all critical sets.

In the algorithm that follows, we write Deadline$_A$(I_h), instead of Deadline(I_h). This is because the algorithm follows a lazy update policy on this variable. In each iteration, a single deadline is updated (in step 8), instead of many. In the algorithm, leeways and criticality are to be computed in terms of the values of Deadline$_A$( ), rather than in terms of Deadline( ). Also in the algorithm, an interval is marked if and only if it has been laid down; all intervals are initially unmarked.

Algorithm for k-layout

```
1. Deadline_A($I_r$) = $n$ for all $r$.
2. For $i = 1$ to $n$ do
3.     Let $h_0$ be the minimum $h$ such that $I_h$ is critical.
4.     Let $z$ be the unmarked interval in $I_{h_0}$ with minimum right coordinate.
5.     Layout($i$) = $z$; Mark $z$.
6.     Let $r$ be the last index such that $z$ is a neighbor of L_array($r$).
7.     /* Among unmarked intervals $y$, $y$ is a neighbor of $z$ if and only if $y$ = L_array($r'$) for some $r' \leq r$. */
8.     Deadline_A($I_r$) = min($i + k$, Deadline_A($I_r$)).
9.     If the leeway of $I_r$ is negative, then quit(failure: bandwidth of $G$ exceeds $k$).
```

The major difference between this algorithm and the algorithm of [8] is the replacement of the sets S_j^q in [8] by the equivalent notions of deadlines and criticality. If our algorithm were to update all deadlines in each iteration, it would be clearly equivalent to the algorithm of [8], and the proof of correctness of [8] would serve also as proof of correctness of this algorithm. In the next paragraph, we explain why the lazy updating policy of this algorithm is valid.

That updating a single deadline in each iteration is sufficient is a consequence of the following considerations. Suppose that $r' < r$ and that Deadline($I_{r'}$) = Deadline(I_r). (1) If some interval in $I_r - I_{r'}$ is not laid down, then Leeway($I_{r'}$) $>$ Leeway(I_r), so the deadline of $I_{r'}$ is superfluous. (2) In the contrary case, all intervals in $I_r - I_{r'}$ are laid down, so Leeway($I_{r'}$) $=$ Leeway(I_r); however, the set of not-laid-down intervals in $I_{r'}$ equals the corresponding subset of I_r. So, if $I_{r'}$ is the smallest critical set, it does not matter if the algorithm treats I_r as the smallest one instead. Where $I_{r'}$ is the smallest critical set $r \geq r'$ and all intervals of $I_r - I_{r'}$ are laid down, we call I_r a *virtually smallest* critical set.

Given arrays LR_array and R_array, step 6 can be computed in $O(1)$ time for the following reasons. Let r be the last index such that z is a neighbor of L_array(r). Then r is the number of intervals u satisfying $u_L \leq z_R$. The number of such intervals equals the index of z_R in LR_array minus the index of z_R in R_array.

In the remainder of the paper, we show that this algorithm can be executed in $O(n \log n)$ time. Since step 6 is not a challenge, the only concerns are steps 3, 4, and 9.

To accomplish this running time, in the next two sections, we design data structures so that each iteration can be performed in $O(\log n)$ time. This necessitates the following:

(A) A data structure to find, in $O(\log n)$ time, a virtually smallest critical set I_{h_0} (step 3);

(B) A data structure to find, in $O(\log n)$ time, the unmarked interval z in I_{h_0} with minimum right coordinate (step 4);

(C) A data structure to determine, in $O(\log n)$ time, if the leeway of I_r is negative (step 9).

A data structure for task (B) is described in §3, and a data structure for tasks (A) and (C) in the following section. Step 8 is accomplished as part of updating the data structure for tasks (A) and (C); a separate array for deadlines is unnecessary.

For tasks (A) and (C), we must know, in each iteration, the first interval in L_array that is not yet laid down. This is easily done: In the initialization of the algorithm, set up a doubly linked list corresponding to L_array and, for each interval z, create a pointer from z to its node in the linked list. Then, in each iteration, in $O(1)$ time, delete from the linked list the interval that is being laid down. In each iteration, the first interval that is not yet laid down can be found, given the header of the linked list, in $O(1)$ time.

3. Selecting the next interval. In this section, we state a computational geometry problem and develop a data structure and algorithm to solve it. In this problem, we are given a set S of points and an on-line sequence of numbers and we compute a permutation of S satisfying certain conditions (or determine that no such permutation exists). At the end of the section, we then show how task (B) from the previous section can be reduced to the computational geometry problem.

Let S be a set of n points in the real plane. To simplify the exposition, we presume that no two points have the same x-coordinate or y-coordinate. Let $x_1, x_2, \ldots, x_n$ be an on-line sequence of numbers—each x_i arrives in *iteration i*. Each x_i is interpreted as an x-coordinate. The objective, in each iteration i, is either to select a point from S (to be called s_i) or to give an error signal. In particular, call the set of points in $S - \{s_1, s_2, \ldots, s_{i-1}\}$ that are left of x-coordinate x_i the *candidate set* for s_i. If the candidate set for s_i is nonempty, s_i is chosen as the member of the candidate set having minimum y-coordinate. However, if the candidate set for s_i is empty, an error signal is returned instead.

Define a *balanced binary tree* of depth d to be a binary tree of depth d such that, when all nodes at depth d are removed, a complete binary tree remains.

The algorithm to be developed for this problem uses a data structure based on the segment tree of Bentley [2], [12, p. 13]. The data structure is a balanced binary tree T with n leaves. Leaves of T correspond to points of S, in order of x-coordinate: Leaf α is left of leaf β if and only if the point corresponding to α is left of the point corresponding to β. Where point s_i corresponds to leaf β, we sometimes refer to β as leaf s_i. The y-coordinate of s_i is then regarded as the y-coordinate of the leaf. Each node γ of T contains a pointer to the leaf in its subtree that has minimum y-coordinate; we let min_leaf(γ) be that leaf.

Initializing T is straightforward. Establishing the correspondence between leaves and points of S takes $O(n)$ time if points in S are already sorted by x-coordinate, and $O(n \log n)$ time otherwise. Computing min_leaf can be done in $O(n)$ time, for it takes $O(1)$ time to compute min_leaf(γ) if min_leaf is already computed for the children of γ.

In iteration i, we must select s_i or determine that the candidate set is empty and we update the data structure to assure that s_i is never again selected. We discuss updating the data structure first. To assure that s_i is not again selected, change its y_coordinate to ∞. This requires recomputing min_leaf(β) for each node β in the path from leaf s_i to the root. This can be done in $O(\log n)$ time.

To select s_i, we first determine the leftmost leaf that is not left of x_i or (alternatively) we determine that all leaves are left of x_i. This can be done in $O(\log n)$ time by a binary search on leaves. If all points are left of x_i, the root points to the point in $S - \{s_i, s_2, \dots, s_{i-1}\}$ having minimum y-coordinate. Suppose instead that α is the leftmost leaf that is not left of x_i and let P be the path from α to the root. Let Q be the set of nodes β such that β is the left child of a node γ, and γ and its right child are both in P. Then Q is the set of nodes in Bentley's segment tree that represent the interval of points that are left of x_i. We can determine Q in $O(\log n)$ time and determine s_i in $O(\log n)$ time by examining the $|Q|$ y-coordinates associated with $\{\text{min\_leaf}(\beta) : \beta \in Q\}$. However, if the point computed for s_i has its y-coordinate ∞, then the candidate set for s_i is empty, and an error signal should be returned instead of a member of S.

Thus T can be initialized in $O(n \log n)$ time, and each iteration can be executed in $O(\log n)$ time. Then initialization plus n iterations take a total of $O(n \log n)$ time. (We note that, where points in S are not initially sorted, this is an optimal algorithm in any model for which sort is an $\Omega(n \log n)$ problem, since sort can be reduced to this problem.)

Task (B) of the previous section can be reduced to this problem as follows. Let I be the set of n intervals. Define S as $\{(z_L, z_R) : z \in I\}$. In iteration t of the algorithm of the previous section the task is to find the unmarked interval in I_{h_0} with minimum right coordinate. Now an interval is unmarked if and only if it is not yet laid down, which, in turn, occurs if and only if it was not selected in previous iterations. Translated into terms of the point set S, the objective is to find the point in S that was not selected in previous iterations and among the leftmost h_0 leaves has minimum y-coordinate. This problem is not harder than the problem we already solved, since, given h_0, we can in $O(1)$ time determine an x-coordinate x_0 so that exactly h_0 leaves are left of x_0. (Actually, it is an easier problem—the binary search mentioned above can be discarded.) We note that in task (B) the critical interval I_{h_0} is known to contain at least one interval that is not laid down, so, as applied to task (B), the algorithm of this section will always return a point, not an error signal.

4. Finding the first critical set. In this section, we describe a data structure that maintains leeways of the sets I_h. Its principal purpose is to enable us, in $O(\log n)$ time, to find a virtually smallest critical set I_h, without getting misled by those I_h that are already fully laid down. The data structure also enables us, in step 9 of the algorithm, to detect if the leeway of a set is negative. States of this data structure are displayed for the interval model of Fig. 2.

The path from a leaf σ to the root is written as P_σ. We define the *left cut from* P_σ to be the set of arcs $\alpha\gamma$ of the tree where α is a left child of γ, and P_σ contains γ but not α. If the arcs of the left cut from P_σ were to be removed from the tree, the leaves that would

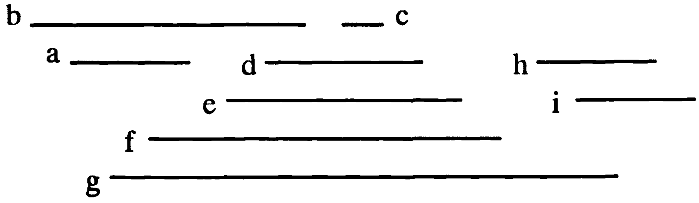


FIG. 2. *Model of interval graph, of bandwidth* 4.

no longer be in the same component with the root would be precisely the leaves properly to the left of σ.

The data structure is a balanced binary tree, like the segment tree of the previous section; we additionally require that each nonleaf have two children. (See Fig. 3.) Like it, leaves represent intervals, in order of the left ends of the intervals: Leaf h of the tree (counting from the leftmost leaf of the tree) corresponds to the interval whose left end is at L_array(h). We speak of leaf h as *laid down* if the interval it represents is laid down. Leaves play a dual role: In addition to representing a single interval, leaf h also represents the set I_h.

We define weights on the arcs of the tree as follows. Let γ be a node, whose left child is α and right child is β. The weight of arc $\beta\gamma$ is 0. The weight of arc $\alpha\gamma$ is the number of already laid down leaves that are descendents of β. A result of this definition is that the number of intervals outside of I_h that are already laid down equals the sum of weights of the arcs on P_h. Then, for a set I_h that is not fully laid down, Leeway(I_h) = Deadline(I_h) $-$ h $-$ weight(P_h), where the weight of P_h is computed as the sum of weights of its arcs.

Recall from §2 that, if I_h is fully laid down, the leeway of I_h is infinite. To accomplish this, weights of some arcs are set to $-\infty$, as follows. Suppose that I_h is the largest fully-laid-down set. Place a weight of $-\infty$ on all arcs of the left cut from P_m, where $m = h + 1$. Now the leeway of each fully laid down set I_h is infinite, and the leeway of each I_h that is not fully laid down is as described in the previous paragraph.

We define values on nodes of the tree as follows. For the leaf I_h, value(I_h) = Deadline$_A$(I_h) $-$ h. Then, at moments when Deadline$_A$(I_h) has been updated so that it equals Deadline(I_h), the value of leaf I_h equals the leeway of I_h, under the optimistic

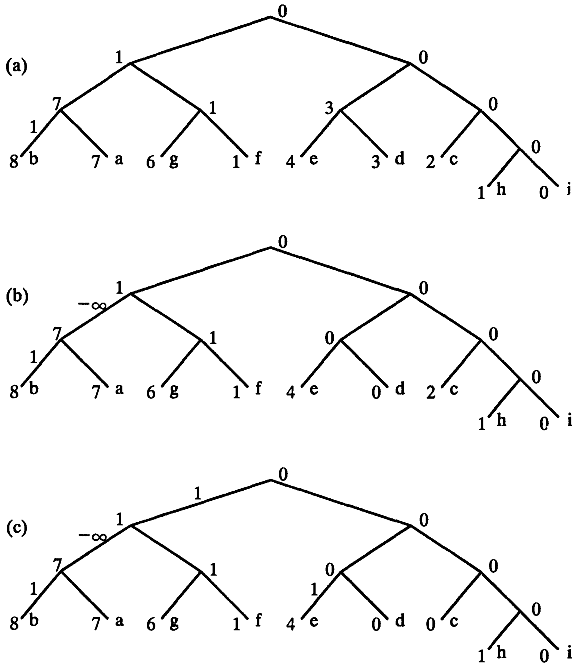


FIG. 3. *Balanced binary tree for the interval graph of Fig. 2. The interval corresponding to each leaf is indicated. Nonzero weights on arcs are shown, as are values on nodes.* (a), (b), *and* (c) *display quantities at the end of iterations* 1, 2, *and* 3, *respectively.*

assumption that no interval outside of I_h has been laid down. For nonleaf γ having left child α and right child β, the value at γ is set to min (value(α) − weight($\alpha\gamma$), value(β)). Then the value at γ equals the minimum leeway of sets I_h whose leaves are descendents of γ, under the optimistic assumption that no interval to the right of this subtree has been laid down (at moments when $\text{Deadline}_A(I_h) = \text{Deadline}(I_h)$, for the appropriate h). Also, γ has a pointer to the child that determined value(γ); in the case of a tie, the pointer is to point to α rather than β.

This tree can be initialized in $O(n)$ time. The weight of each arc is zero. For each leaf h, $\text{Deadline}_A(h) = n$, so the value of the leaf is initialized to $n - h$; the value of each nonleaf may be computed in $O(1)$ time, given the values of its children.

In the remainder of this section we denote by z_h the interval corresponding to leaf h.

In each iteration of the algorithm, this tree can be updated in $O(\log n)$ time. Let interval z_h be laid down in iteration t. Let r be the last index such that z_h is a neighbor of L_array(r) (so r is the last leaf that z_h is a neighbor of). Suppose first that I_h does not become fully laid down in this iteration. To update weights on arcs, we increment the weight of arcs in the left cut from P_h, which may be done in $O(\log n)$ time. To update the value of leaf r, we execute the assignment value(r) = min(value(r), $t + k - r$) (which corresponds to $\text{Deadline}_A(r) = \min(\text{Deadline}_A(r), t + k)$). Values of all other leaves are left unchanged. The only nonleaves whose value must be updated are those on either P_h or P_r. Next, suppose that I_h becomes fully laid down. To update the tree, first find h', so that $z_{h'}$ is the first not-yet-laid-down interval (as described at the end of §2) and make the weight of all arcs in the left cut from $P_{h'}$ be $-\infty$. Then update the value of leaf r and update the values of nodes that are on either P_r or $P_{h'}$.

Given this tree, we can detect (in step 9) if I_r has negative leeway in $O(1)$ time by checking if the value at the root is negative. This is because the value at the root is the minimum leeway of sets I_h, and, if any set has negative leeway, I_r does. Where no set has negative leeway, we can find a virtually smallest critical set by following the value pointers, starting from the root. Thus tasks (C) and (A) can be accomplished in $O(\log n)$ time.

Figure 3 displays this data structure for the interval graph of Fig. 2. When the algorithm is executed with $k = 4$, interval a is laid down in iteration 1. Figure 3(a) displays the data structure at the end of iteration 1: Since the deadline of I_4 becomes 5, the value of the fourth leaf is set to 1. The left cut from P_2 contains a single arc, whose value becomes 1. In iteration 2, b is laid down. Since I_2 becomes fully laid down, the value of arcs on the left cut from P_3 are set to $-\infty$ (see Fig. 3(b)). Also, the deadline of I_6 becomes 6, so the value of the sixth leaf is set to 0: I_6 is critical. In iteration 3, the interval laid down is d (d is selected by the data structure of the previous section as the remaining interval in I_6 having minimum right end). (See Fig. 3(c).) The weights of the two arcs in the left cut from P_6 are incremented. Both I_4 and I_7 become critical (the former because of the incrementation of the weight of an arc, and the latter because the deadline of I_7 becomes 7). In iteration 4, the interval to be laid down is selected from I_4. The layout constructed by the algorithm is a, b, d, f, g, e, c, h, i.

5. Conclusion. The $O(n \log n)$ algorithm presented in this paper leads to several open problems.

Many problems that are NP-complete on graphs in general have $O(n)$ algorithms on interval graphs, where n is the number of intervals in the interval model. Some such problems are clique cover, coloring, maximum clique, maximum independent set, and minimum dominating set of an interval graph [6], [7]. For these algorithms, the ends of intervals are assumed presorted. With this paper, bandwidth joins Hamiltonian cycle as problems for which $O(n \log n)$ algorithms are known [10] but no $O(n)$ algorithm. Are these $O(n \log n)$ algorithms optimal when ends of intervals are assumed presorted? Resolution of this question seems difficult.

The algorithm of this paper enables the bandwidth of an interval graph to be computed in $O(n \log^2 n)$ time. To do this, we perform a binary search on values of k and, for each such value of k, run the k-bandwidth algorithm. Can this complexity be brought down?

A proof of correctness of the algorithm is presented in [8]. That proof shows that there is no minimal graph G for which the algorithm fails. Is there a more direct proof of correctness of the strategy underlying the algorithm—a proof that, given an interval model and a layout L of bandwidth k but which differs from the layout produced by this algorithm, L may be adjusted to a layout L', also of bandwidth at most k, and which is "closer" to following the strategy than L is? Such a proof would likely involve choosing a class of transformations such that the set of all layouts of bandwidth at most k is connected by those transformations. An analogous result was proved in [13]: The class of Hamiltonian cycles of an interval graph is connected under the transformation of edge exchange. The method of such a proof may be strong enough to show that, if Layout$(1 \cdots t)$ is preassigned, with Layout(1) being the interval with leftmost right end, there exists an extension of this partial layout to a full layout of bandwidth k if and only if the strategy behind this algorithm successfully completes the layout.

REFERENCES

[1] S. F. Assman, G. W. Peck, M. M. Syslo, and J. Zak, *The bandwidth of caterpillars with hairs of length* 1 *and* 2, SIAM J. Algebraic Discrete Meth., 2 (1981), pp. 387–393.

[2] J. L. Bentley, *Algorithms for Klee's Rectangle Problems*, Department of Computer Science, Carnegie-Mellon Univ., Pittsburgh, PA, 1977.

[3] A. A. Bertossi and A. Gori, *Total domination and irredundance in weighted interval graphs*, SIAM J. Discrete Math., 1 (1988), pp. 317–327.

[4] K. S. Booth and G. S. Lueker, *Testing for the consecutive ones property, interval graphs and graph planarity using PQ-tree algorithms*, J. Comput. Syst. Sci., 13 (1976), pp. 335–379.

[5] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[6] U. I. Gupta, D. T. Lee, and J. Y.-T. Leung, *Efficient algorithms for interval graphs and circular-arc graphs*, Networks, 12 (1982), pp. 459–467.

[7] S. K. Kim, *Optimal parallel algorithms on sorted intervals*, In Proc. 27th Annual Allerton Conf. on Comm., Control, and Computing, Monticello, IL, 1989, Univ. of Illinois Press, pp. 766–775.

[8] D. J. Kleitman and R. V. Vohra, *Computing the bandwidth of interval graphs*, SIAM J. Discrete Math., 3 (1990), pp. 373–375.

[9] D. Kratsch, *Finding the minimum bandwidth of an interval graph*, Inform. Comput., 74 (1987), pp. 140–158.

[10] G. K. Manacher, T. A. Mancus, and C. D. Smith, *An optimum* $\theta(n \log n)$ *algorithm for finding a canonical Hamiltonian path and a canonical Hamiltonian circuit in a set of intervals*, Inform. Process. Lett., 35 (1990), pp. 205–211.

[11] B. Monien, *The bandwidth minimization problem for caterpillars with hair length* 3 *is NP-complete*. SIAM J. Algebraic Discrete Meth., 7 (1986), pp. 505–512.

[12] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*, Springer-Verlag, Berlin, New York, 1988.

[13] A. P. Sprague, *Edge exchanges on Hamiltonian cycles in interval graphs*, Congr. Numer., 85 (1991), pp. 111–122.

SIAM J. Discrete Math.
Vol. 7, No. 2, pp. 221-229, May 1994

THE LAPLACIAN SPECTRUM OF A GRAPH II*

ROBERT GRONE[†] AND RUSSELL MERRIS[‡]

Abstract. Let G be a graph. Denote by $D(G)$ the diagonal matrix of its vertex degrees and by $A(G)$ its adjacency matrix. Then $L(G) = D(G) - A(G)$ is the Laplacian matrix of G. The first section of this paper is devoted to properties of Laplacian integral graphs, those for which the Laplacian spectrum consists entirely of integers. The second section relates the degree sequence and the Laplacian spectrum through majorization. The third section introduces the notion of a d-cluster, using it to bound the multiplicity of d in the spectrum of $L(G)$.

Key words. degree sequence, majorization, Laplacian integral

AMS subject classification. 05C50

1. Laplacian integral graphs.

Let $G = (V, E)$ be a graph with vertex set $V = V(G) = \{v_1, v_2, \ldots, v_n\}$ and edge set $E = E(G)$. Denote the degree of vertex v_i by $d(v_i)$. Let $D(G) = \text{diag}\,(d(v_1), d(v_2), \ldots, d(v_n))$ be the diagonal matrix of vertex degrees. The *Laplacian matrix* is $L(G) = D(G) - A(G)$, where $A(G)$ is the (0, 1)-adjacency matrix. It follows from the Geršgorin disc theorem that $L(G)$ is positive semidefinite and from the matrix-tree theorem (or from [3]) that its rank is $n - w(G)$, where $w(G)$ is the number of connected components of G. (More on the Laplacian may be found in [10] or [17].) Denote the spectrum of $L(G)$ by

$$S(G) = (\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = 0$ are the eigenvalues of $L(G)$. If more than one graph is involved, we may write $\lambda_i(G)$ in place of λ_i. The multiplicity of λ as an eigenvalue of $L(G)$ will be denoted $m_G(\lambda)$.

The central theme of this article is the occurrence of integers in $S(G)$. If the spectrum consists entirely of integers, we say G is *Laplacian integral*. The study of graphs whc e *adjacency* spectra consist entirely of integers was begun in [11]. Cvetković [4] proved that the set of connected, r-regular, adjacency integral graphs is finite. When $r = 2$, there are three such graphs, namely, C_3, C_4, and C_6. (When $r = 3$, there are 13 such graphs [2], [21].) If G is r-regular, then λ is an eigenvalue of $L(G)$ if and only if $r - \lambda$ is an eigenvalue of $A(G)$. Thus, the theory of Laplacian integral graphs coincides with its adjacency counterpart on regular graphs. Elsewhere, there can be remarkable differences. Consider, for example, the 112 connected graphs on six vertices. Six of them are adjacency integral. Of these six, five are regular: C_6 and its complement, K_6, $K_{3,3}$, and the cocktail party graph. The sixth is the tree obtained by joining the centers of two copies of P_3 with a new edge. As we have observed, the first five of these are also Laplacian integral; the sixth is not. On the other hand, there are a total of 37 connected Laplacian integral graphs on six vertices [18].

One general difference between the two theories concerns complements. Let J_n be the n-by-n matrix, each of whose entries is 1. Then, for any graph G on n vertices,

* Received by the editors November 20, 1991; accepted for publication (in revised form) March 23, 1993.

[†] Department of Mathematical Sciences, San Diego State University, San Diego, California 92182. The work of this author was supported by National Science Foundation grant DMS9007048.

[‡] Department of Mathematics and Computer Science, California State University, Hayward, California 94542. The work of this author was supported by a California State University Research, Scholarship and Creative Activity award and by National Security Agency grant MDA904-90-H-4024.

$L(G) + L(G^c) = nI_n - J_n$. It follows that the eigenvalues of $L(G^c)$ are

$$(1) \qquad \lambda_i(G^c) = n - \lambda_{n-i}(G), \quad 1 \leq i < n, \quad \text{and} \quad 0.$$

In particular, G is Laplacian integral if and only if G^c is Laplacian integral, and $\lambda_1(G) \leq n$, with $m_G(n) = w(G^c) - 1$. Another difference involves trees. While some interesting work has been done on adjacency integral trees [11], [22], a complete description seems unlikely in the near future. On the other hand, T is Laplacian integral if and only if $T = K_{1,n-1}$ [7], [15, Cor. 2].

Let $G = (V, E)$ be a graph with n vertices and m edges. If $e = \{u, v\} \in E$ and $w \notin V$, then e is said to be *subdivided* when it is replaced by $\{u, w\}$ and $\{w, v\}$. Of course, replacing the edge replaces the graph; the new graph has $n + 1$ vertices and $m + 1$ edges.

Example 1. Denote by G_n the graph obtained from K_{n-1}, $n > 2$, by subdividing one edge; then G_n is Laplacian integral. This is because the complement of G_n is a graph with two connected components, one isomorphic to K_2 and the other to $K_{1,n-3}$. The same result does not hold for the adjacency matrix; $A(G_5)$ has three irrational eigenvalues. If two edges of K_3 are subdivided, the result is C_5, which is neither adjacency nor Laplacian integral.

If every edge of G is subdivided, the resulting graph $s(G)$ has $n + m$ vertices and $2m$ edges. It is called the *subdivision* of G. (*Note*. $S(G)$ is an n-tuple; $s(G)$ is a graph.)

THEOREM 1. *Let G be a connected, r-regular, Laplacian integral graph on n vertices. Then $s(G)$ is Laplacian integral if and only if $G = K_n$.*

Proof. The result is trivial for $r < 2$. If $r = 2$, then (as we have seen above) $G = C_3$, C_4, or C_6, Of these, $s(C_3) = C_6$ is Laplacian integral, whereas $s(C_4) = C_8$ and $s(C_6) = C_{12}$ are not. Thus, we may assume that $r \geq 3$ (so $n \geq 4$). Let m denote the number of edges in G. It was shown in [13] (see [17]) that

$$\det(xI_{m+n} - L(s(G))) = (-1)^n(x - 2)^{m-n} \det(x(r + 2 - x)I_n - L(G)).$$

Therefore, α is an eigenvalue of $L(s(G))$ if and only if $\alpha = 2$ or $\alpha(r + 2 - \alpha)$ is an eigenvalue of $L(G)$. If $G = K_n$, then $r = n - 1$, and the eigenvalues of $L(s(G))$ satisfy $m_{s(G)}(0) = 1$, $m_{s(G)}(1) = n - 1$, $m_{s(G)}(2) = n(n - 3)/2$, $m_{s(G)}(n) = n - 1$, and $m_{s(G)}(n + 1) = 1$.

Conversely, the eigenvalues of $L(G)$ are all of the form $\lambda = \alpha(r + 2 - \alpha)$. Since $r + 1$ is the minimum value taken by this product when both factors are constrained to be positive integers, we deduce that $\lambda_{n-1}(G) \geq r + 1$. Thus, the trace of $L(G)$ is at least $(n - 1)(r + 1)$. However, trace $L(G) = rn$. Combining these, we conclude that $r \geq n - 1$. □

It is proved in [11, Cor. 6] that the line graph of a regular adjacency integral graph is adjacency integral. Since the line graph of a regular graph is regular, this result carries over to the Laplacian case. So, for example, the Petersen graph is Laplacian integral since it is the complement of the line graph of K_5. Recall that a graph is (r, s)-*semiregular* if it is bipartite with a bipartition (V_1, V_2) in which each vertex of V_1 has degree r and each vertex of V_2 has degree s. The next result was proved by Mohar [17, Thm. 3.9].

PROPOSITION 1. *Let G be a connected, (r, s)-semiregular, Laplacian integral graph. Then its line graph is Laplacian integral.*

COROLLARY 1. *The line graph of the subdivision of K_n is Laplacian integral.*

Ultimately, we would like to "explain" all the integral graphs (whether Laplacian or adjacency). Harary and Schwenk [11] identified various families of adjacency integral graphs. Inevitably, they found graphs that belong to none of them. One such graph is the line graph of the subdivision of K_4. It is clear from Corollary 1 that this graph does, in fact, belong to a natural family. However, because the characteristic polynomial of

$A(s(K_4))$ is $x^2(x^2 - 2)^3(x^2 - 6)$, this fact is not immediately evident from the adjacency perspective.

Let G and H be graphs on disjoint sets of vertices. Their *union* $G + H$ is the graph with vertex set $V(G + H) = V(G) \cup V(H)$ and edge set $E(G + H) = E(G) \cup E(H)$. The *join* of G and H may be defined by $G \vee H = (G^c + H^c)^c$. It is the graph obtained from $G + H$ by adding new edges from each vertex of G to every vertex of H. Clearly, the union and join of Laplacian integral graphs are Laplacian integral.

The *product* of graphs G and H is the graph $G \times H$, whose vertex set is the Cartesian product $V(G) \times V(H)$. Suppose that $v_1, v_2 \in V(G)$ and $u_1, u_2 \in V(H)$. Then (v_1, u_1) and (v_2, u_2) are adjacent in $G \times H$ if and only if one of the following conditions is satisfied: (i) $v_1 = v_2$ and $\{u_1, u_2\} \in E(H)$ or (ii) $\{v_1, v_2\} \in E(G)$ and $u_1 = u_2$. For example, the line graph of $K_{p,q}$ is $K_p \times K_q$. If G and H are Laplacian integral graphs, then $G \times H$ is Laplacian integral [17, Thm. 3.5]. Let $G^2 = G \times G$. The subgraph of G^2 induced on $W = \{(v_i, v_j): i < j\} \subset V(G^2)$ is called $G^{[2]}$ [8]. Since both $(x^2 - 7x + 8)^2$ and $(x^2 - 10x + 20)$ are factors of the characteristic polynomial of $L(C_6^{[2]})$, $G^{[2]}$ does not generally propagate Laplacian integrality. Still, this construction is the source of numerous Laplacian integral graphs.

2. Majorization and the degree sequence. Recall that a nonincreasing sequence $(d) = (d_1, d_2, \ldots, d_n)$ of positive integers is said to be *graphic* if there exists a (simple) graph having degree sequence (d). The theory of graphic sequences has a rich tradition nicely summarized in [19].[1] Another perspective on the discussion of §1 would be to approach nonincreasing sequences $(\lambda) = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ of nonnegative real numbers in a similar way; i.e., what are necessary and/or sufficient conditions for (λ) to be the spectrum of the Laplacian matrix of a (simple) graph, and what special conditions arise if we require (λ) to be an integer sequence? One condition follows immediately from a theorem of Schur. It involves the property of majorization. Suppose that (a) and (b) are finite nonincreasing sequences of real numbers. Then (a) is said to *majorize* (b) if

$$a_1 \geq b_1,$$

$$a_1 + a_2 \geq b_1 + b_2,$$

$$a_1 + a_2 + a_3 \geq b_1 + b_2 + b_3,$$

and so on, with equality at the end; that is, the sum of all the a's is equal to the sum of all the b's.

In the introductory paragraph, we defined $D(G)$ to be the diagonal matrix of vertex degrees. We abuse that language now by letting $D(G)$ also denote the *sequence* of vertex degrees in nonincreasing order, i.e., $D(G) = (d_1, d_2, \ldots, d_n)$. (We do not assume that $d_i = d(v_i)$.) Since the spectrum of any symmetric, positive semidefinite matrix majorizes its main diagonal [20], [14, p. 218], the following is immediate from the definitions.

PROPOSITION 2. *Let G be a graph. Then $S(G)$ majorizes $D(G)$.*

The most immediate consequence of Proposition 2 is the inequality $\lambda_1 \geq d_1$. In fact, this inequality is subject to some improvement, as we now show.

THEOREM 2. *Suppose that $G = (V, E)$ is a connected a graph with $n > 2$ vertices. If $S = \{u_1, u_2, \ldots, u_k\} \subset V$, let $G[S] = (S, E[S])$ be the subgraph of G induced on S. Suppose that $E[S]$ consists of r pairwise disjoint edges. Then*

$$\lambda_1 + \cdots + \lambda_k \geq d(u_1) + \cdots + d(u_k) + k - r.$$

(Recall that $d(u_i)$ need not equal d_i.)

[1] Unfortunately, [19] contains several annoying misprints.

The inequality in Theorem 2 suggests it might be useful to define

$$\Sigma_k(G) = \max\left\{\sum_{i=1}^{k} d(u_i) \colon \{u_1, \ldots, u_k\} \text{ independent in } G\right\}.$$

Curiously enough, the analogous quantity

$$\sigma_k(G) = \min\left\{\sum_{i=1}^{k} d(u_i) \colon \{u_1, \ldots, u_k\} \text{ independent in } G\right\}$$

is useful in the study of Hamiltonian cycles [9].

COROLLARY 2. *If G has an edge, then*

$$\lambda_1 \geq d_1 + 1. \tag{2}$$

Proof. If $d_1 = 1$, then every connected component of G is either an isolated vertex or a copy of K_2. In this case, $\lambda_1 = 2$ and $d_1 = 1$. If $d_1 > 1$, let w be a vertex of G of degree d_1. Let C be a connected component of G containing w. Then we may apply Theorem 2 to the subgraph C with $k = 1$ and $u_1 = w$. Since $L(C)$ is a direct summand of $L(G)$, $\lambda_1(G) \geq \lambda_1(C)$. □

Corollary 2 improves the earlier result [7, Thm. 3.7] $\lambda_1 \geq d_1 + d_1/(n - 1)$. (It can be shown that inequality in (2) is strict whenever $d_1 < n - 1$.)

COROLLARY 3. *Let G be a connected graph on $n > 2$ vertices. Then $\lambda_1 + \lambda_2 \geq d_1 + d_2 + 1$. If there are two nonadjacent vertices in G having degrees d_1 and d_2, then $\lambda_1 + \lambda_2 \geq d_1 + d_2 + 2$.*

Proof. The proof is immediate from Theorem 2. Either $r = 1$ or $r = 0$. □

Example 2. Let G be the graph shown in Fig. 1. Then $S(G) = (5.1, 3.8, 1.5, 1, 0.6, 0)$ and $\lambda_1 + \lambda_2 < d_1 + d_2 + 2$.

CONJECTURE 1. *Let G be a graph with $m \geq 1$ edges. Then* $S(G)$ *majorizes* the sequence $(d_1 + 1, d_2, d_3, \ldots, d_{n-1}, d_n - 1)$.

Proof of Theorem 2. Order the m edges in E arbitrarily. For each edge $e = \{v, w\}$, designate one of v, w to be the "positive end" of e and the other to be the "negative end." Thus, G is given a fixed but arbitrary *orientation.* If $e \in E$ and $v \in V$, define $s(v, e) = +1$ if v is the positive end of e, -1 if it is the negative end, and 0 otherwise. The vertex-edge *incidence matrix* afforded by the orientation is the n-by-m matrix $Q = Q(G)$ whose (v, e)-entry is $s(v, e)$. It turns out that $L(G) = QQ^t$, independently of the orientation. The matrix $K(G) = Q^tQ$ depends on the orientation for the signs of its off-diagonal entries. In any event, $K(G)$ and $L(G)$ share the same nonzero eigenvalues.

Suppose that x is some real m-tuple. It is convenient to think of its components as indexed by E, so the "vth" component" of Qx is

$$\sum_{e \in E} s(v, e)x_e.$$

Therefore,

$$x^tK(G)x = \sum_{v \in V}\left[\sum_{e \in E} s(v, e)x_e\right]^2. \tag{3}$$

Suppose that $E[S] = \{e_1, e_2, \ldots, e_r\}$. Without loss of generality, we may assume $e_i = \{u_i, u_{k-i+1}\}$ and $d(u_i) \leq d(u_{k-i+1})$, $1 \leq i \leq r$. Choose an orientation of $E(G)$ so that u_i is the positive end of each of the $d(u_i)$ edges incident with it, $1 \leq i \leq k - r$. In addition, we may prescribe that u_i is the positive end of each of the $d(u_i) - 1 > 0$ edges,

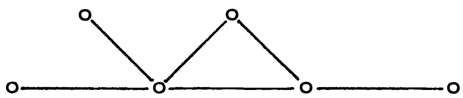


FIG. 1

other than e_{k-i+1}, incident with it, $k - r < i \leq k$. For each $u \in S$, define a real m-tuple $x(u)$ as follows: $x(u)_e$, the coordinate of $x(u)$ corresponding to the edge $e \in E$, is 1 if u is the positive end of e, and 0 otherwise. Then $\{x(u)\colon u \in S\}$ is an orthogonal set of vectors. Moreover, $\|x(u_i)\|^2 = d(u_i)$ if $i \leq k - r$ and $d(u_i) - 1$ if $i > k - r$. Let $y(u) = x(u)/\|x(u)\|$, $u \in S$. Then, for $1 \leq i \leq k - r$,

$$\sum_{e \in E} s(v, e)y(u_i)_e = \begin{cases} d(u_i)^{1/2} & \text{if } v = u_i, \\ -d(u_i)^{-1/2} & \text{if } \{v, u_i\} \in E, \\ 0 \quad \text{otherwise,} \end{cases}$$

and, for $i > k - r$,

$$\sum_{e \in E} s(v, e)y(u_i)_e = \begin{cases} (d(u_i) - 1)^{1/2} & \text{if } v = u_i, \\ -(d(u_i) - 1)^{-1/2} & \text{if } \{v, u_i\} \in E \backslash E[S], \\ 0 \quad \text{otherwise.} \end{cases}$$

So, from (3),

$$y(u_i)^t K(G) y(u_i) = \begin{cases} d(u_i) + 1, & i \leq k - r, \\ d(u_i), & i > k - r. \end{cases}$$

Since $\{y(u_i)\colon 1 \leq i \leq k\}$ is an orthonormal set of vectors,

$$\begin{aligned} \sum_{i=1}^{k} \lambda_i &\geq \sum_{i=1}^{k} y(u_i)^t K(G) y(u_i) \\ &= \sum_{i=1}^{k} d(u_i) + k - r. \qquad \square \end{aligned}$$

Nonincreasing integer sequences are frequently pictured by means of so-called Ferrers–Sylvester diagrams. The diagram for $(d) = (5, 5, 5, 4, 4, 4, 3)$ is pictured on the left in Fig. 2. Its *transpose* is the diagram pictured on the right corresponding to the conjugate partition $(d)^t = (7, 7, 7, 6, 3)$. In general, the *conjugate* of a nonincreasing integer sequence

$$(a) = (a_1, a_2, \ldots, a_p)$$

is

$$(a)^t = (a_1^t, a_2^t, \ldots, a_q^t),$$

where $q = a_1$ and a_i^t is the number of elements in the set $\{j\colon a_j \geq i\}$.

PROPOSITION 3. *Let (d) be a graphic sequence. Then $(d)^t$ majorizes (d).*

Proof. Suppose that $d_i \geq i$, $1 \leq i \leq k$, and $d_{k+1} < k + 1$. (In Fig. 2, $k = 4$.) Define $r_i = d_i + 1$, $1 \leq i \leq k$ and $r_i = d_i$, $k < i \leq n$. Ryser's necessary and sufficient condition for (d) to be graphic is that $(r)^t$ majorize (r) (see [1, §11.5]). Note that $r_i^t = d_i^t$, $1 \leq i \leq k$

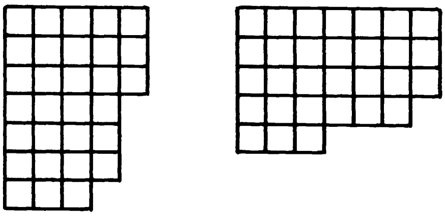


FIG. 2. *Ferrers–Sylvester diagrams.*

and $r_i^t = d_i^t + c_i$, $i > k$, where the c_i are nonnegative integers *that add up to k*. So, for $1 \le j \le k$,

$$\sum_{i=1}^{j} d_i^t = \sum_{i=1}^{j} r_i^t \ge \sum_{i=1}^{j} r_i = \sum_{i=1}^{j} (d_i + 1) = j + \sum_{i=1}^{j} d_i > \sum_{i=1}^{j} d_i.$$

For $j > k$, let $x_j = c_1 + c_2 + \cdots + c_{j-k}$. Then

$$x_j + \sum_{i=1}^{j} d_i^t = \sum_{i=1}^{j} r_i^t \ge \sum_{i=1}^{j} r_i = k + \sum_{i=1}^{j} d_i.$$

So, since $x_j \le k$,

$$\sum_{i=1}^{j} d_i^t \ge (k - x_j) + \sum_{i=1}^{j} d_i \ge \sum_{i=1}^{j} d_i. \qquad \square$$

Propositions 2 and 3 raise the natural question of whether $S(G)$ and $D(G)^t$ are majorization comparable. (An infinite family of "maximal" graphs for which $D(G)^t = S(G)$ is discussed in [16].)

CONJECTURE 2. *Let G be a connected graph. Then $D(G)^t$ majorizes $S(G)$.*

If Conjecture 2 is true, then

$$\lambda_{n-1} \ge d_{n-1}^t; \tag{4}$$

i.e., the number of vertices of G of degree $n - 1$ is bounded above by λ_{n-1}, the "algebraic connectivity" of G [7]. To verify (4), suppose that G is a graph with k vertices of degree $n - 1$. If $k = n$, then $G = K_n$ and $\lambda_{n-1} = n$. Otherwise, G^c has at least $k + 1$ components, the largest of which has at most $n - k$ vertices. So, $\lambda_1(G^c) \le n - k$. Thus, $\lambda_{n-1}(G) = n - \lambda_1(G^c) \ge k$.

If $D(G)^t$ majorizes $S(G)$ for connected graphs, then $D(G)^t$ majorizes $S(G)$ for all graphs as can easily be seen, e.g., by the condition of Hardy, Littlewood, and Polya [14, p. 22]. So, we may as well restrict our attention to connected graphs, where the number of vertices of G having degree at least 1 is $d_1^t = n$. The inequality $n \ge \lambda_1$ is clear from (1). Indeed, if G is a connected r-regular graph, then $D(G) = (r, r, \ldots, r)$ and $D(G)^t = (n, n, \ldots, n)$ majorizes $S(G)$.

A second step toward proving Conjecture 2 would be to show

$$d_1^t + d_2^t \ge \lambda_1 + \lambda_2. \tag{5}$$

We are not able to establish (5) in all cases. However, some partial results along these lines will emerge from the following result.

THEOREM 3. *Let G be a connected graph and suppose that w is a cut vertex of G. If the largest component of $G - w$ contains r vertices, then $r + 1 \ge \lambda_2(G)$.*

Proof. Denote by $L(w)$ the $(n - 1)$-by-$(n - 1)$ principal submatrix of $L(G)$ obtained by striking out the row and column corresponding to w. Let α be the largest eigenvalue of $L(w)$. By the Cauchy interlacing inequalities, $\lambda_1 \ge \alpha \ge \lambda_2$.

If $C_1, C_2, \ldots, C_k$ are the connected components of $G - w$, let S_i be the union of $\{w\}$ and the vertices of C_i, $1 \leq i \leq k$. Let $G_i = G[S_i]$ be the subgraph of G induced by S_i and write $L_i = L(G_i)$, $1 \leq i \leq k$. Then $L(w)$ is the direct sum of $L_i(w)$, $1 \leq i \leq k$, where $L_i(w)$ is the principal submatrix of L_i obtained by striking out the row and column corresponding to w. It follows that α is the largest eigenvalue of $L_i(w)$ for some i. By another application of the Cauchy inequalities, we conclude that $\alpha \leq \lambda_1(G_i)$ for some i. By (1) and the hypotheses, $r + 1 \geq \lambda_1(G_i)$ for all i. □

A pendant *neighbor* is a vertex adjacent to a vertex of degree 1.

COROLLARY 4. *Let G be a connected graph with $n > 2$ vertices. Suppose that w is a pendant neighbor of G adjacent to k pendant vertices. Then $n - k \geq \lambda_2$.*

Proof. If $G = K_{1,n-1}$, then $n - k = 1 = \lambda_2$. Otherwise, the largest component of $G - w$ contains at most $n - k - 1$ vertices, so the result is immediate from Theorem 3. □

We now return to (5). Since $n = d_1^t \geq \lambda_1$, it would suffice to show that $d_2^t \geq \lambda_2$. Now d_2^t is the number of vertices of G having degree at least 2. That is, $d_2^t = n - p$, where p is the number of pendant vertices. However, as we now see, it is not generally true, even for trees, that $n - p \geq \lambda_2$.

Example 3. Let T be the tree on six vertices obtained by joining the centers of two copies of P_3 by a new edge. Then $D(T) = (3, 3, 1, 1, 1, 1)$ and $D(T)^t = (6, 2, 2)$. To one decimal place, $S(T) = (4.6, 3, 1, 1, 0.4, 0)$. (Recall that T is the only adjacency integral graph on six vertices that is not Laplacian integral.) Here $n = 6$, $p = 4$, and $n - p = 2 < 3 = \lambda_2(T)$.

COROLLARY 5. *Let G be a connected, Laplacian integral graph with n vertices, p of which are pendants. Then $n - p \geq \lambda_2$.*

Proof. Suppose that G has a total of q pendant neighbors altogether. It is proved in [10, Thm. 3.11] that the number of eigenvalues of $L(G)$, multiplicities included, lying in the open interval $[0, 1)$ is at least q. Since we are assuming that G is connected, $m_G(0) = 1$. Thus, $q > 1$ contradicts the hypothesis that G is Laplacian integral. We conclude that all p pendants of G share the same neighbor, so the result follows from Corollary 4. □

3. Eigenvalues and graph structure. In a natural way, the majorization questions of §2 have led to the relationship between the Laplacian spectrum and graph structure. We proceed to develop more results along these lines, beginning with a relative of Corollary 5.

PROPOSITION 4. *Let G be a graph with $n > 2$ vertices, p of which are pendants. If $\lambda_1 = n$, then all p of the pendants are adjacent to the same neighbor w, $d(w) = n - 1$, and $\lambda_2 \leq n - p$. (In particular, if T is a tree, then $\lambda_1(T) = n$ if and only if $T = K_{1,n-1}$.)*

Proof. From (1), $\lambda_1(G) = n$ if and only if G^c is disconnected. If G had two distinct pendant neighbors, then G^c would be connected. So, there is a unique pendant neighbor w. We conclude from Corollary 4 that $\lambda_2 \leq n - p$. If $d(w)$ were less than $n - 1$ then, again, G^c would be connected. □

PROPOSITION 5. *Let G be a connected graph. Let $P = \{v_1, v_2, \ldots, v_k\}$ be a set of pendant vertices of G, all of which are adjacent to the same neighbor w. Suppose that $\lambda \neq 1$. If (the multiplicity) $m_G(\lambda) > 1$, then $m_{G-P}(\lambda) > 0$. (In particular, $\lambda \leq n - k$.) Moreover, if $m_{G-P}(\lambda) > 1$, then $m_G(\lambda) > 0$.*

Proof. Let $w = v_{k+1}$. If λ is a multiple eigenvalue of $L(G)$, it has an eigenvector x whose $(k + 1)$st component is 0. Then $L(G)x = \lambda x$ forces $x_1 = x_2 = \cdots = x_k = 0$. It follows that $y = (0, x_{k+2}, \ldots, x_n)$ is an eigenvector of $L(G - P)$, affording λ. Because $G - P$ has $n - k$ vertices, $\lambda \leq n - k$. To obtain the final assertion, let w be the first vertex of $G - P$. If λ is a multiple eigenvalue of $L(G - P)$, then it is afforded by an eigenvector

y whose first component is 0. Let x be the vector obtained from y be inserting k zeros at the beginning. Then x is an eigenvector of $L(G)$, affording λ. □

PROPOSITION 6. *Let G be a graph with n vertices and $k \geq 1$ spanning trees. If λ is a positive integer eigenvalue of $L(G)$, then $\lambda \mid nk$. If G is Laplacian integral, then $\lambda^t \mid nk$, where $t = m_G(\lambda)$, the multiplicity of λ as an eigenvalue of $L(G)$.*

Proof. Let $p(x)$ be the characteristic polynomial of $L(G)$. Since G is connected, $L(G)$ has rank $n - 1$, so we may factor $p(x)$ as $xf(x)$. (The polynomial $f(x)$ has been called the "T-polynomial" of G [5], [12].) As $f(0)$ is the coefficient of x in $p(x)$, it is the sum of the determinants of the $(n - 1)$-by-$(n - 1)$ principal submatrices of $L(G)$. By the matrix-tree theorem, each of these n determinants has the value k. Thus, $f(0) = nk$. Since $f(x)$ is a monic polynomial with integer coefficients, $f(\lambda) = 0$ if and only if $\lambda \mid f(0)$. On the other hand, $f(0)$ is the $(n - 1)$th elementary symmetric function of the eigenvalues of $L(G)$, i.e., $f(0)$ is the product of the nonzero eigenvalues of $L(G)$. So, if G is Laplacian integral, $\lambda^t \mid f(0)$. □

DEFINITION. *Let G be a graph. A* cluster *of G is an independent set of two or more vertices of G, each of which has the same set of neighbors. (The set of neighbors of vertex v is $\{u \in V\colon \{u, v\} \in E\}$.) The* degree *of a cluster is the cardinality of its shared set of neighbors, i.e., the common degree of each vertex in the cluster. A d-*cluster *is a cluster of degree d. The number of vertices in a d-cluster is its* order. *A collection of two or more d-clusters is* independent *if the sets of vertices comprising the d-clusters are pairwise disjoint. (The neighbor sets of independent d-clusters need not be disjoint.)*

The next result extends the work of Faria on the "star degree" of a graph [6].

THEOREM 4. *Let G be a graph with k independent d-clusters of orders $r_1, r_2, \ldots, r_k$. Then $m_G(d) \geq r_1 + r_2 + \cdots + r_k - k$.*

Example 4. The graphs G_1 and G_2 in Fig. 3 are the smallest pair of nonisomorphic, connected, Laplacian integral, Laplacian cospectral graphs. They share the spectrum $S(G_1) = S(G_2) = (7, 6, 6, 4, 4, 3, 0)$. Both pictures are drawn so that the top row of vertices is a 4-cluster: G_1 contains a 4-cluster of order 2, while G_2 contains one of order 3. With $d = 4$ and $k = 1$, Theorem 4 asserts that $m_G(4) \geq 1$ for $G = G_1$ and $m_G(4) \geq 2$ for $G = G_2$; the bound is sharp for G_2 but not for G_1. On the other hand, an examination of the spectrum shows there is no point in looking for a 5-cluster in either graph.

Proof of Theorem 4. The independent d-clusters correspond to k nonoverlapping principal submatrices of $L(G)$. Each submatrix is d-times an r_i-by-r_i identity matrix, $i = 1, 2, \ldots, k$. Suppose that one of these principal submatrices includes rows and columns numbered s and t in $L(G)$, $s < t$. (That is, suppose that v_s and v_t belong to the same d-cluster in G.) Let x be the column vector with $x_i = 1$ if $i = s$, -1 if $i = t$, and 0, otherwise. Because v_s and v_t belong to the same d-cluster, they have the same neighbors. Hence, $L(G)x = dx$. A d-cluster of order r_i affords $r_i - 1$ linearly independent eigenvectors of this type, and eigenvectors of this type arising from independent clusters are linearly independent. □

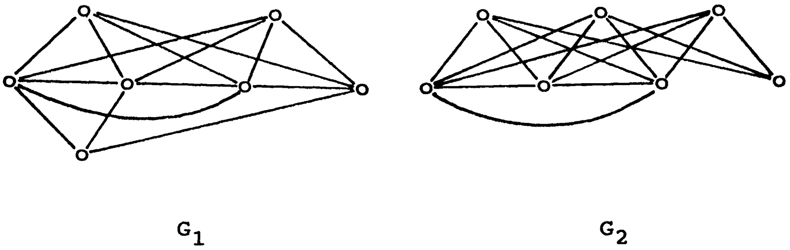


FIG. 3

COROLLARY 6. *Let G be a graph with an r-clique, $r \geq 2$. Suppose that every vertex of the clique has the same set of neighbors outside the clique. Let the degree of each vertex of the clique be d, so $d - r + 1$ is the number of vertices not belonging to the clique but adjacent to every member of the clique. Then $m_G(d + 1) \geq (r - 1)$.*

Proof. The clique becomes an $(n - d - 1)$-cluster of G^c of order r. □

Example 5. Let G be the graph G_2 of Fig. 3. The three vertices of G of degree 5 are a 3-clique satisfying the hypotheses of Corollary 6. Hence $m_G(6) \geq 2$, and, again, we find that G_2 affords a sharp bound.

Note added in proof. Theorem 4 was first proved by Isabel Faria, who communicated it to Merris long before the present paper was contemplated. Unfortunately, Merris filed it away and forgot about it. Fortunately, the result will appear under Faria's name in the article, "Multiplicity of Integer Roots of Polynomials of Graphs," to be published by *Linear Algebra Appl.*

REFERENCES

[1] J. A. BONDY AND U. S. R. MURTY, *Graph Theory With Applications*, North-Holland, New York, 1976.

[2] F. C. BUSSEMAKER AND D. M. CVETKOVIĆ, *There are exactly* 13 *connected, cubic, integral graphs*, Univ. Beograd Publ. Elektrotehn. Fak. Ser. Mat. Fiz., 544–576 (1976), pp. 43–48.

[3] I. M. CHAKRAVARTI, *On a characterization of irreducibility of a nonnegative matrix*, Linear Algebra Appl., 10 (1975), pp. 103–109.

[4] D. M. CVETKOVIĆ, *Cubic integral graphs*, Univ. Beograd Publ. Elektrotehn. Fak. Ser. Mat. Fiz. 498–541 (1975), pp. 107–113.

[5] E. A. DINIC, A. K. KELMANS, AND M. A. ZAITSEV, *Nonisomorphic trees with the same T-polynomial*, Inform. Process. Lett., 6 (1977), pp. 73–76.

[6] I. Faria, *Permanental roots and the star degree of a graph*, Linear Algebra Appl., 64 (1985), pp. 255–265.

[7] M. FIEDLER, *Algebraic Connectivity of graphs*, Czech. Math. J., 23 (98) (1973), pp. 298–305.

[8] ———, *Irreducibility of compound matrices*, Comment. Math. Univ. Carolin., 20 (1979), pp. 737–743.

[9] R. J. GOULD, *Updating the hamiltonian problem—a survey*, J. Graph Theory, 15 (1991), pp. 121–157.

[10] R. GRONE, R. MERRIS, AND V. S. SUNDER, *The Laplacian spectrum of a graph*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 218–238.

[11] F. HARARY AND A. J. SCHWENK, *Which graphs have integral spectra*?, in Graphs and Combinatorics, R. A. Bari and F. Harary, eds., Lecture Notes in Math 406, Springer-Verlag, Berlin, 1974.

[12] Y. HATTORI, *Nonisomorphic graphs with the same T-polynomial*, Inform. Process. Lett., 22 (1986), pp. 133–134.

[13] A. K. KEL'MANS, *Properties of the characteristic polynomial of a graph*, Kibernetiky-na službu kommunizmu, v. 4, Energija, Moskva-Leningrad, 1967, pp. 27–41. (In Russian.)

[14] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[15] R. MERRIS, *Characteristic vertices of trees*, Linear Multilinear Algebra, 22 (1987), pp. 115–131.

[16] ———, *Degree maximal graphs are Laplacian integral*, Linear Algebra Appl., to appear.

[17] B. MOHAR, *The Laplacian spectrum of graphs*, Graph Theory, Combinatorics, and Applications, Y. Alavi et al., eds., John Wiley, New York, 1991, pp. 871–898.

[18] D. L. POWERS, *Graph Eigenvectors*, unpublished tables.

[19] G. SIERKSMA AND H. HOOGEVEEN, *Seven criteria for integer sequences being graphic*, J. Graph Theory, 15 (1991), pp. 223–231.

[20] I. SCHUR, *Über eine Klasse von Mittelbildungen mit Anwendungen die Determinanten*, Theorie Sitzungsber. Berlin. Math. Gesellschaft, 22 (1923), pp. 9–20.

[21] A. J. SCHWENK, *Exactly thirteen connected cubic graphs have integral spectra*, in Theory and Applications of Graphs, Y. Alavi et al., eds., Springer-Verlag, Berlin, 1978, pp. 516–533.

[22] M. WATANABE AND A. J. SCHWENK, *Integral starlike trees*, J. Austral. Math. Soc. Ser. A, 28 (1979), pp. 120–128.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 230–237, May 1994

009

A $2d - 1$ LOWER BOUND FOR TWO-LAYER KNOCK-KNEE CHANNEL ROUTING*

TOM LEIGHTON†

Abstract. This paper describes a two-point net channel routing problem with density d that requires channel width $2d - 1$ in the two-layer knock-knee channel routing model. This means that the $(2d - 1)$-track algorithms of Rivest, Baratz, and Miller [1981 CMU *Conference on* VLSI *Systems and Computations*, Oct. 1981, pp. 153–159], Bolognesi and Brown [unpublished manuscript, Coordinated Science Laboratory, University of Illinois at Urbanna-Champaign, 1982], Frank [*Combinatorica*, 2 (1982), pp. 361–371], Mehlhorn, Preparata, and Sarrafzadeh [University of Saarbrucken Tech. Rep., Saarbrucken, Germany, Nov. 1984], and Berger et al. [*J. Assoc. Comput. Mach.*, 1994, to appear], are, in some cases, optimal. Thus, any improvement of these algorithms must rely on problem features other than density (such as flux [*Advances in Computing Research* 2 (VLSI *Theory*), F. P. Preparata, ed., JAI Press, Greenwich, CT, 1984, pp. 205–229]) or must make fundamental changes in the wiring model (such as increasing the number of layers [*IEEE Trans. Comput.*, C-33 (1984), pp. 427–437] or allowing wires to overlap [see Berger et al., above], [*Algorithmica*, 1 (1986), pp. 223–232]).

Key words. channel routing, VLSI, lower bounds

AMS subject classifications. 68Q25, 68Q35, 68R99

1. Introduction. Channel routing plays a central role in the development of automated layout systems for integrated circuits. Many layout systems first place modules on a chip or circuit board and then wire together terminals on different modules that should be electrically connected. This wiring problem is often solved by heuristically partitioning the given space into rectangular channels and then assigning to each such channel a set of wires that are to pass through it. This solution reduces a "global" wiring problem to a set of disjoint (and hopefully easier) "local" channel routing problems. For this reason, channel routing problems have been intensively studied for over a decade, and numerous heuristics and approximation algorithms have been proposed.

In most channel routing problems, the *channel* consists of a rectilinear grid of *tracks* (or rows) and *columns*. Along the top and bottom tracks are *terminals*, and terminals with the same label form a *net*. A net with r terminals is called an *r-point net*. The smallest net is a *two-point* net. If $r > 2$, we have a *multipoint* net. The channel routing problem is to connect all the terminals in each net using horizontal and vertical wires that are routed along the underlying rectilinear grid. The goal is to complete the wiring using the minimum number of tracks, i.e., to minimize the *width* of the channel. Often, no constraint is placed on the number of columns used at either end of the channel.

Wire segments are physically located on one of the one or more *layers*. Wire segments in different layers can be connected with *contact cuts*, which can be thought of as very short wire segments running through grid points perpendicular to the routing surface. No two wires can change layers in the same space without being connected.

A variety of models have been proposed for channel routing, with differences depending on the number of layers allowed and on the ways in which wires are allowed to interact. One of the most popular models is the knock-knee model proposed by Rivest,

* Received by the editors November 30, 1988; accepted for publication (in revised form) March 25, 1993. This research was supported by U.S. Air Force contract F49620-92-J-0125, Defense Advanced Research Projects Agency contracts N00014-91-J-1698 and N00014-92-J-1799, and a National Science Foundation Presidential Young Investigator Award with matching funds from AT&T and IBM.

† Mathematics Department and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Baratz, and Miller [8] and Thompson [9]. In the knock-knee model, wires are allowed to cross or share corners (i.e., knock-knees) but are not allowed to overlap for any distance. In this paper, we focus on two-layer knock-knee channel routing problems with only two-point top-to-bottom nets. More formally, we define a *net* $N_i = (p_i, q_i)$ to be an ordered pair of integers specifying an *entry column* p_i (at the *top* of the channel) and an *exit column* q_i (at the *bottom* of the channel). A net is said to be *rising* if $q_i < p_i$, *falling* if $p_i < q_i$, and *trivial* if $p_i = q_i$. A *channel routing problem* is a collection of n nets such that no two nets have a common entry column or a common exit column. A *solution* to a channel routing problem consists of an integer t and a collection of n wires $W_1, \ldots, W_n$ such that W_i enters the grid in the p_ith column of the zeroth row and exits the grid in the q_ith column of the $(t + 1)$th row. Routing is allowed in rows 1 through t of the grid and in any number of columns. Although wires may alternate layers (via contact cuts), they cannot overlap except at crossover or jog points. The *width* of a solution is t, the number of horizontal tracks used to route the wires. The *optimal width* of a channel routing problem is simply the smallest value of t for which there is a solution.

Many algorithms have been discovered for solving two-layer knock-knee channel routing problems with width $2d - 1$, where d is the density of the channel routing problem. (For example, see [2], [3], [4], [6], and [8].) The *density* of a channel routing problem is the maximum over all $x \in \mathcal{R}$ of the number of nets crossing the vertical cut of the channel at x. A net $N_i = (p_i, q_i)$ is said to *cross* the cut of the channel at x if $p_i < x < q_i$ or $q_i < x < p_i$. For example, the channel routing problem displayed in Fig. 1 has density 3.

It is easy to see that the density of a problem is a lower bound on the channel width needed for its solution. Hence, the $(2d - 1)$-width algorithms of [2], [3], [4], [6], [8] are within a factor of two times optimal for any problem. As many practical channel routing problems have solutions with width very close to d, however, it was hoped that the $(2d - 1)$-width algorithms could be improved (e.g., to produce $3d/2$-width solutions for problems with density d).

In this paper, we show that such an improvement is not possible. In particular, we construct a two-point net channel routing problem with density d (for any d) that requires channel width $2d - 1$. Hence, the $(2d - 1)$-width algorithms are optimal in the worst case and cannot be improved by density considerations alone.

It is worth pointing out that improvements in the performance of the algorithms can be achieved if additional parameters are considered or if different models are used. For example, Baker, Bhatt, and Leighton [1] showed that any two-point net channel routing problem with density d and flux f can be solved using $d + O(f)$ tracks, even if

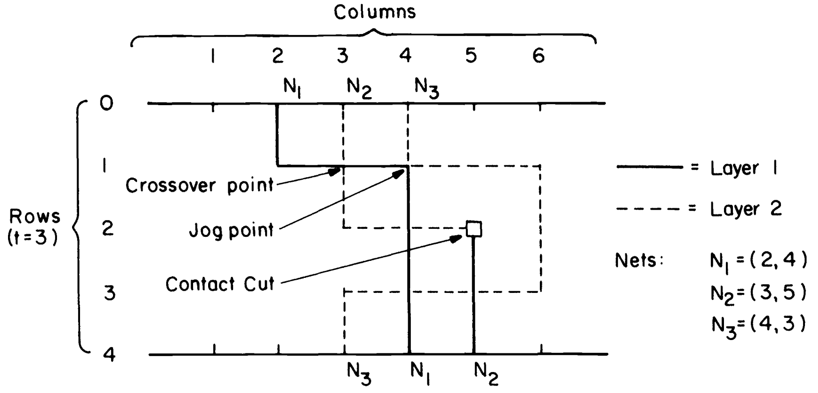


FIG. 1. *A solution to a channel routing problem using the knock-knee model.*

knock-knees are not allowed. (The definition of *flux* is somewhat technical, but roughly corresponds to a horizontal measure of density. In the worst case, $f \sim \sqrt{n}$, but f is often a small constant in practical problems.) Alternatively, Preparata and Lipski [7] showed that d tracks are sufficient if three layers of wiring are allowed, thus achieving the lower bound for every problem. More recently, Berger et al. [2] showed that $d + O(\sqrt{d})$ tracks are sufficient using two layers if unit-length vertical wire segments are allowed to overlap other vertical wire segments. (For a more complete listing of other bounds and algorithms for channel routing, refer to [2].)

The paper is divided into sections as follows. In §2 we construct the density-d lower bound example and in §3 we prove that it requires $2d - 1$ tracks. We conclude with some remarks.

2. The lower bound construction. Although it may seem somewhat complicated at first, our lower bound construction is quite simple. The problem consists primarily of trivial nets along with a few falling nets arranged as widely spaced 1-shifts. For example, an informal illustration of the construction for $d = 2$ is shown in Fig. 2.

More formally, our hard density-d channel routing problem $R_{d,n}$ is composed of nets $N_1, N_2, \ldots, N_n$, where $N_i = (i, q_i)$, and

$$q_i = \begin{cases} i + 4^s s!^2 & \text{if } i \equiv 4^{s-1} \bmod 4^s s!^2 \text{ for some } s \le d, \\ i & \text{otherwise} \end{cases}$$

for $1 \le i \le n$. In what follows, we use $n = d^2 4^{d+1} d!^2$, although $R_{d,n}$ can be defined for any n such that $n \equiv 0 \bmod 4^d d!^2$.

We first show that $R_{d,n}$ is a well-defined channel routing problem. To do this, we must show that q_i is well defined and that no pair of nets share an entry or exit column. Assume for the purposes of contradiction that q_i is not well defined for some i. Then there are two integers $s > r$ such that $i \equiv 4^{s-1} \bmod 4^s s!^2$ and $i \equiv 4^{r-1} \bmod 4^r r!^2$. The first congruence implies that 4^{s-1} divides i and thus that 4^r divides i (since $s - 1 \ge r$). The latter fact clearly contradicts the second congruence. Thus, q_i is well defined for each i.

By definition, it is clear that no two nets have the same entry column. In what follows, we show that no two nets have the same exit column. If two nets were to share the same exit column, then there would be two integers $i < j$ such that $q_i = q_j$. Suppose (for the purposes of contradiction) that this is the case. Then, since $q_j \ge j$, it is clear that $q_i > i$ and thus that $q_i = i + 4^s s!^2$, where $i \equiv 4^{s-1} \bmod 4^s s!^2$ for some s. Thus $q_j = q_i \equiv 4^{s-1} \bmod 4^s s!^2$. If $j = q_j$, then $j = 4^{s-1} \bmod 4^s s!^2$, and $q_j = j + 4^s s!^2$, which is a contradiction.

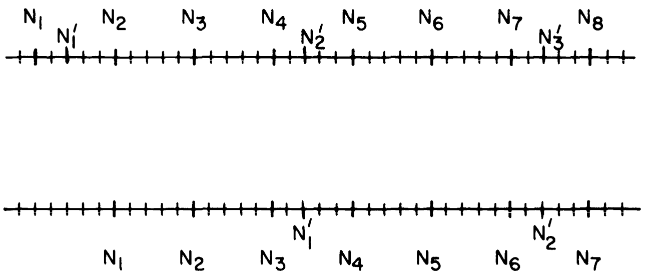


FIG. 2. *An informal drawing of the lower bound construction for $d = 2$. The problem consists mostly of trivial nets (which are not shown), along with a widely spaced 1-shift sequence of nets ($N_1, N_2, N_3, \ldots$) and another even more widely spaced 1-shift sequence of nets ($N'_1, N'_2, \ldots$).*

This means that $q_j > j$ and thus that $q_j = j + 4^r r!^2$, where $j \equiv 4^{r-1} \bmod 4^r r!^2$ for some r. Thus $q_j \equiv 4^{r-1} \bmod 4^r r!^2$. Since $q_i = q_j$, we can conclude by the arguments of the previous paragraph that $r = s$ and thus that $i = j$, a contradiction.

It is also easy to show that $R_{d,n}$ has density d. This is due to the fact that the nontrivial nets of $R_{d,n}$ can be partitioned into blocks $B_1, \dots, B_d$, where

$$B_s = \{(4^{s-1}, 4^{s-1} + 4^s s!^2), (4^{s-1} + 4^s s!^2, 4^{s-1} + 2 \cdot 4^s s!^2), \dots,$$
$$(4^{s-1} + n - 4^s s!^2, 4^{s-1} + n)\}$$

for $1 \le s \le d$. Each vertical cut of the channel is crossed by at most one net from each block, and thus the density of $R_{d,n}$ is at most d. In fact, the density is precisely d, since any cut in the range $4^{d-1} < x < n + 1$ is crossed by a net from every block.

Note that the nets in any block B_s correspond to a 1-shift that is spread out across the channel. Hence, our lower bound construction consists only of a collection of interlaced 1-shifts and trivial nets.

3. Proof of the lower bound. The proof that $R_{d,n}$ requires channel width $2d - 1$ proceeds in two parts. In Part 1, §3.1, we show that wires passing through a large region of the channel must be routed in a highly restricted manner and that there is such a region that contains the solution to a subproblem of $R_{d,n}$, which is isomorphic to $R_{d,f(d)}$, where $f(d) = 4^d d!^2$. In Part 2, §3.2, we show by induction on d that any solution of $R_{d,f(d)}$ in such a region has width at least $2d - 1$.

3.1. Part 1. Assume for the purposes of contradiction that $R_{d,n}$ has a solution with width t, where $t < 2d - 1$. Given any integer y such that $0 \le y \le t$, let P_y be the path that travels between columns n and $n + 1$ past rows 0 through y, then between rows y and $y + 1$ past columns n through 1, and finally between columns 1 and 0 past rows $y + 1$ to $t + 1$. For example, see Fig. 3.

Precisely n nets must cross P_y at least once for any y. In what follows, we are concerned primarily with the *initial crossing* of each wire on its route from the ith column of row 0 to the q_ith column of row $t + 1$. Since at most t wires can cross the vertical portions of P_y, at least $n - t$ wires must initially cross P_y on its horizontal portion. (Note that wires at such points are travelling in a *downward direction* across P_y.) Since n columns of the grid cross the horizontal portion of P_y, at most t of them can fail to contain an initial crossing of P_y. Summing over all paths P_y, we find that at most $t(t + 1) \le 4d^2 - 6d + 2 < 4d^2$ *unit segments* of columns 1 through n (those between row y and row $y + 1$ for some y) fail to contain a wire that is travelling in a downward direction.

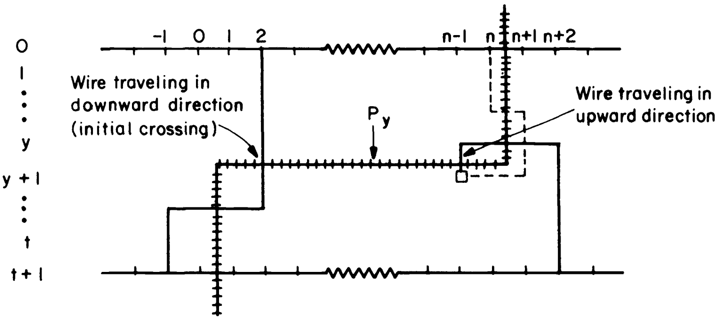

0
1
y
y + 1
t
t + 1
-1 0 1 2
n-1 n n+1 n+2
Wire traveling in downward direction (initial crossing)
P_y
Wire traveling in upward direction

FIG. 3. *The path P_y.*

Partition columns $1, \ldots, n$ into $4d^2$ groups, each with $(n/4d^2) = 4^d d!^2$ consecutive columns. By the previous argument, we know that, in at least one of these groups (say the mth group), every column segment contains a wire travelling in the *downward* direction. In other words, there is an m ($1 \le m \le 4d^2$) such that every segment in columns $(m - 1)4^d d!^2 + 1$ through $m4^d d!^2$ contains a wire travelling in the downward direction. Henceforth, we refer to these columns as the *restricted region* of the channel.

Note that the ith column of the restricted region is the same as column $(m - 1)4^d d!^2 + i$ of the original problem. Since $4^d d!^2 \equiv 0 \bmod 4^s s!^2$ for all $s \le d$, we know that, for all $s \le d$,

$$i \equiv 4^{s-1} \bmod 4^s s!^2 \quad \text{iff}\, (m - 1)4^d d!^2 + i \equiv 4^{s-1} \bmod 4^s s!^2.$$

Hence, the columns in the restricted region contain a subproblem that is isomorphic to $R_{d,f(d)}$, where we define $f(d) = 4^d d!^2$.

3.2. Part 2. We now show that any solution to $R_{d,f(d)}$ in a restricted region (i.e., a region in which every vertical segment of the first $f(d)$ columns contains a wire travelling in the downward direction) requires $2d - 1$ tracks. The proof is by induction on d. This hypothesis is certainly true for $d = 1$. In what follows, we assume the hypothesis is true for $d - 1$ to verify it for d. First, however, we must establish some useful facts concerning routings in restricted regions. For instance, each row in such a region is either *full* (i.e., each unit segment of the row contains a wire) or *empty* (none of the unit segments contains a wire). This simple but powerful observation follows from the fact that every column in a restricted region is full. Thus a row that is neither full nor empty must contain a point that is incident to two-column wire segments and one-row wire segment. This is clearly impossible.

It is natural to group the rows together into (alternating) blocks of full rows and empty rows. Such blocks are said to be *full* or *empty*, respectively. Note that a wire can change layers only in an empty block and can change columns only in a full block. For example, see Fig. 4.

Only a restricted set of column changes is possible in a full block of rows. For example, consider a full block that contains a long horizontal wire segment. Since we are routing in a restricted region, the columns crossing the wire must contain wires as in Fig. 5(a). Since all of the crossing wires are moving in a downward direction, no two can be connected by a unit horizontal wire, and thus the rows above and below the long horizontal wire also contain long horizontal wires (provided that they are not empty, of course). For example, see Fig. 5(b). By repeating this argument for the remaining rows, we can deduce that all the rows of the full block contain long horizontal wires. For example, see Fig. 5(c).

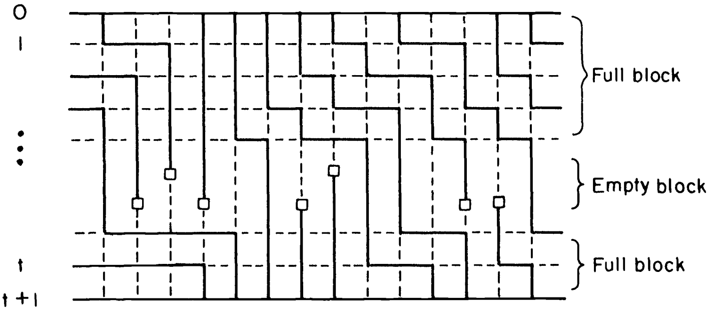


FIG. 4. *Wiring in a restricted region.*

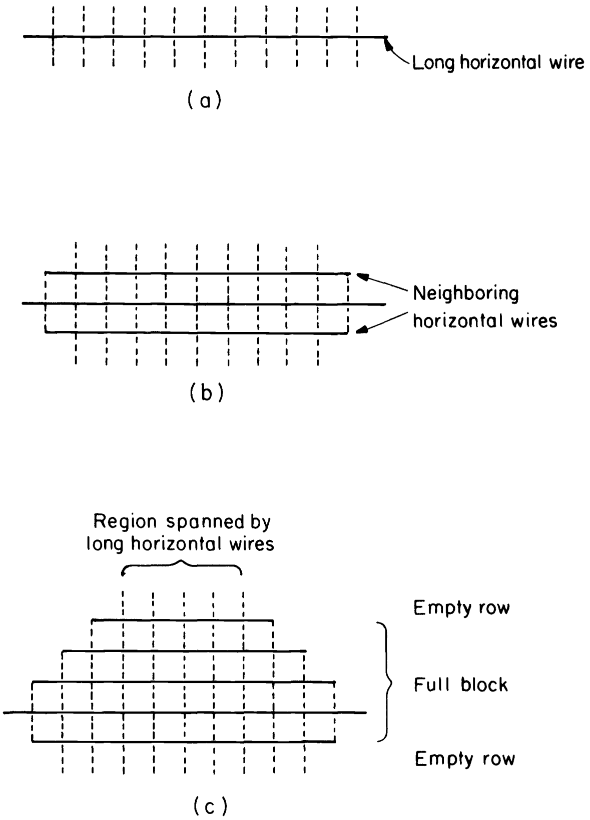

Long horizontal wire
(a)
Neighboring horizontal wires
(b)
Region spanned by long horizontal wires
Empty row
Full block
Empty row
(c)

FIG. 5. *Impact of a long horizontal wire on a full block.*

We are now ready to prove that any routing of $R_{d,f(d)}$ in a restricted region requires $2d-1$ tracks. For the purposes of contradiction, assume otherwise and consider the net $N_i = (i, q_i)$, where $i = 4^{d-1}$ and $q_i = 4^{d-1} + 4^d d!^2$. Because we are in a restricted region, the corresponding wire in the solution must eventually travel from column 4^{d-1} to column $4^d d!^2$ (at least) in a downward (or level) fashion. This means that some row of the solution contains a horizontal wire segment of length at least $(4^d d!^2 - 4^{d-1})/t$. Thus, the block containing this row resembles that in Fig. 5(c). In particular, there is a subregion consisting of at least

$$\frac{4^d d!^2 - 4^{d-1}}{t} - 2(t-1) \ge \frac{4^d d!^2 - 4^{d-1}}{2d-2} - 2(2d-3)$$

consecutive columns that is spanned by continuous horizontal wires in every row of the block.

Partition the $f(d)$ columns of the restricted region into $f(d)/f(d-1) = 4d^2$ groups, each consisting of $f(d-1) = 4^{d-1}(d-1)!^2$ contiguous columns. By the analysis at the end of §3.1, we can show that each of these groups of columns (except the first) contains a routing problem that is isomorphic to $R_{d-1,f(d-1)}$. This is because the ith column in the mth group $(1 \le m \le 4d^2)$ is column $(m-1)4^{d-1}(d-1)!^2 + i$ in the restricted region, and

$$i \equiv 4^{s-1} \bmod 4^s s!^2 \quad \text{iff}\ (m-1)4^{d-1}(d-1)!^2 + i \equiv 4^{s-1} \bmod 4^s s!^2$$

for $s \leq d - 1$. We must be careful with columns for which

$$(m-1)4^{d-1}(d-1)!^2 + i \equiv 4^{d-1} \bmod 4^d d!^2$$

(since then $q_i = i + 4^{d-1}$ instead of $q_i = i$, as we would want for $R_{d-1,f(d-1)}$), but this only happens for $m = 1$ and $i = 4^{d-1}$ (since $(m-1)4^{d-1}(d-1)!^2 + i \leq 4^d d!^2$), which is why the first group does not contain a routing problem isomorphic to $R_{d-1,f(d-1)}$.

For $d > 1$,

$$\frac{4^d d!^2 - 4^{d-1}}{2d-2} - 2(2d-3) > (2d+2)4^{d-1}(d-1)!^2,$$

and thus the wires of at least $2d$ subproblems isomorphic to $R_{d-1,f(d-1)}$ enter and exit in the columns through which all of the long horizontal wire segments pass. (Note that this rules out both the first and last groups of $f(d-1)$ columns.) Of these, at most $t \leq 2d - 2$ subproblems can have a wire that passes outside those columns (since it must also re-enter on some other row). Thus all the wires of at least one of the subproblems isomorphic to $R_{d-1,f(d-1)}$ are totally contained within the columns spanned by the long horizontal wires. As the wires for this subproblem must pass straight through (downward) the block containing the long wires, the horizontal rows of the block can be removed without affecting the wires for this solution of $R_{d-1,f(d-1)}$. In other words, this means that there is a $(t-1)$-track solution to $R_{d-1,f(d-1)}$ in a restricted region.

In addition, we can also remove an empty track from the region without affecting the solution to $R_{d-1,f(d-1)}$. To see why, first consider the case when the full block with long horizontal wires is preceded and followed by empty blocks. When the full block is removed, we will then have an empty block with at least two rows (one from each of the neighboring blocks). Empty blocks of rows can always be replaced by a single empty row, since wires can only change layers in empty blocks and they only need one row to do so. Hence, we can remove an empty row from the layout without affecting the wiring of $R_{d-1,f(d-1)}$.

If the full block with long horizontal wires is not preceded and followed by empty blocks, then it must be adjacent to the top or bottom track of the channel. In this case, when the full block is removed, we can also remove the single adjacent empty block. This is because we never need an empty block adjacent to the top or bottom of the channel, since we can simply start or end each wire in the desired layer when it enters or exits the channel.

Combining the previous arguments, we find that we can always remove at least one full track and one empty track without affecting the solution to $R_{d-1,f(d-1)}$. This means that there is a solution to $R_{d-1,f(d-1)}$ in a restricted region that uses at most $t - 2 < 2d - 3$ tracks. This contradicts the inductive hypothesis, and thus we know that the solution to $R_{d,f(d)}$ must have used $t \geq 2d - 1$ tracks. This concludes the proof.

4. Remarks. It is interesting to note that the construction consists mostly of trivial nets and does not contain any rising nets whatsoever. A priori, such a channel routing problem might have been considered to be easier than one that corresponds to a *permutation* (i.e., one in which every entry column is also an exit column). As we have seen, however, this is not the case. As we might expect, it is not difficult to modify the construction to prove an identical lower bound for channel routing problems that *are* permutations.

The term "trivial net" is a misnomer. For example, adding trivial nets to some channel routing problems (such as a 2-shift) increases their optimal channel width. In addition, the optimal routing of a trivial net can be quite complicated. For example, the obvious routing of a trivial net $N_i = (p_i, p_i)$ is to run a wire down the p_ith column from row 0 to row $t + 1$. Even if layer changes are allowed for such wires, it is possible to construct examples of $O(d^2)$-net channel routing problems with optimal width $d + O(\log d)$ for which any such solution requires $2d - 1$ horizontal tracks.

Whether the number of wires in our lower bound example can be significantly decreased is an interesting open question. Although the size of our construction was artificially increased to simplify the proof, we do not know of any n-net channel routing problem with density d and optimal width $2d - 1$ for which $n \leq o(d!)$. Thus, it is still possible that a channel routing algorithm could be found that produces solutions with width $d + \log n$.

A closer examination of our lower bound proof reveals that the restricted region of the routing contains at least d full rows and at least $d - 1$ empty rows. This coincides exactly with the behavior of the $(2d - 1)$-track algorithms. Those algorithms use d tracks for routing and $d - 1$ tracks for layer changes. Hence, the need for layer changes is a crucial factor in determining the channel width of problems in the two-layer knock-knee model, and it cannot be ignored.

Acknowledgments. I am grateful to Gary Miller for mentioning this problem to us and for his assistance with this work. I also thank Ron Rivest and the referees for their helpful comments.

REFERENCES

[1] B. Baker, S. N. Bhatt, and T. Leighton, *An approximation algorithm for Manhattan routing*, in Advances in Computing Research 2 (VLSI Theory), F. P. Preparata, ed., JAI Press, Greenwich, CT, 1984, pp. 205–229.

[2] B. Berger, M. Brady, D. J. Brown, and F. T. Leighton, *Nearly optimal bounds and algorithms for multilayer channel routing*, J. Assoc. Comput. Mach., 1994, to appear.

[3] T. Bolognesi and O. Brown, *A Channel Routing Algorithm with Bounded Wire Length*, Coordinated Science Laboratory, University of Illinois, Urbana-Champaign, 1982, unpublished manuscript.

[4] A. Frank, *Disjoint paths in a rectilinear grid*, Combinatorica, 2 (1982), pp. 361–371.

[5] S. Gao and S. Hambrusch, *Two-layer channel routing with vertical unit-length overlap*, Algorithmica, 1 (1986), pp. 223–232.

[6] K. Mehlhorn, F. Preparata, and M. Sarrafzadeh, *Channel Routing in Knock-Knee Mode: Simplified Algorithms and Proofs*, Univ. Saarbrucken Tech. Report, Saarbrucken, Germany, Nov. 1984.

[7] F. P. Preparata and W. Lipski, *Optimal three-layer channel routing*, IEEE Trans. on Comput., C-33 (1984), pp. 427–437.

[8] R. L. Rivest, A. Baratz, and G. Miller, *Provably good channel routing algorithms*, in Proc. 1981 CMU Conf. on VLSI Systems and Computations, Pittsburgh, PA, Oct. 1981, pp. 153–159.

[9] C. Thompson, *A Complexity Theory for* VLSI, Ph.D. thesis, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1980.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 238-244, May 1994

THE DIVISORS OF $x^{2^m} + x$ OF CONSTANT DERIVATIVES AND DEGREE 2^{m-2}*

CLAUDE CARLET†

Abstract. This paper provides a new proof of the fact that the polynomials of degree 2^{m-2} over the Galois field GF(2^m) ($m \geq 2$), which are fully reducible and admit no multiple factor (i.e., which divide $x^{2^m} + x$) and whose derivatives are constant are affine polynomials. The author determines explicitly these polynomials, which are related to a problem in coding theory that is recounted.

Key words. finite field, polynomial, reducible, error correcting code

AMS subject classifications. 12E05, 12E20, 94B15

1. Introduction. Let m be a positive integer, $m \geq 2$ and G the Galois field of order 2^m. Let $F = \text{GF}(2) = \{0, 1\}$ its prime subfield. G may be viewed as an m-dimensional vector space over F.

Little is known about the *developed expressions* of those polynomials over G that divide $x^{2^m} - x$ (that is, $x^{2^m} + x$, since the characteristic is 2). For instance, the question of determining such polynomials of a given degree and of a certain type is not answerable, except in a few cases (cf. [3] in the case of lacunary polynomials). One of these cases is that of those polynomials of degree 2^{m-2} whose derivatives are constant. A subcase (which is, in fact, equivalent) is when the constant term of the polynomial is zero. It is related to a problem in algebraic coding theory (that we recount in §2): the characterization of those elements of the narrow-sense BCH code of length $2^m - 1$ and designed distance $2^{m-2} - 1$ whose weight is equal to that designed distance. Augot, Charpin and Sendrier [1] have recently given a characterization of these elements. Their main result may be stated as follows: Any polynomial of degree 2^{m-2}, whose derivative is a constant, which divides $x^{2^m} + x$ and whose constant term is zero is a linearized polynomial. The proof, which uses Newton identities, is very satisfactory but does not point out significant properties that would explain what leads to such a remarkable result. The purpose of this paper is to exhibit such properties and to deduce a new proof of the result.

We introduce (§3) a property about the divisibility of polynomials over Galois fields of characteristic 2 and deduce (§4) a new proof of the fact that these polynomials (without the condition on the constant term) are all affine (i.e., are sums of linearized polynomials over G and constants). We then determine explicitly the expressions of these polynomials (not done in [1]).

Let us recall that any function from G to G admits a representation as a polynomial over G of degree at most $2^m - 1$. We call it the polynomial expression of the function.

If $P(x)$ is a polynomial over G of any degree, we denote by $P(x) \bmod (x^{2^m} + x)$ the unique polynomial of degree at most $2^m - 1$, which is congruent with $P(x)$ modulo $x^{2^m} + x$.

We denote tr as the trace function from G to F: $\text{tr}(x) = \sum_{i=0}^{m-1} x^{2^i}$. A polynomial $f(x)$ of degree at most $2^m - 1$ is the polynomial expression of a Boolean function (i.e., of a function from G to F) if and only if there exists a polynomial $g(x)$ such that $f(x) = \text{tr}(g(x)) \bmod (x^{2^m} + x)$. The condition is sufficient since tr is a mapping from G to F,

* Received by the editors April 9, 1992; accepted for publication (in revised form) April 12, 1993.

† Institut National de Recherche en Informatique et en Automatique, Bat 10, Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France and Laboratoire de Mathématiques et d'Informatique Fondamentale d'Amiens (LAMIFA), Université de Picardie, France.

and it is necessary since, if f is a Boolean function and a is any element of G such that $\operatorname{tr}(a) = 1$, then we have, for any element u in G, $\operatorname{tr}(af(u)) = f(u) \operatorname{tr}(a) = f(u)$; therefore, according to the unicity of the polynomial expression of f,

$$\operatorname{tr}(af(x)) \bmod (x^{2^m} + x) = f(x).$$

2. Presentation of the problem on BCH codes (cf. [1], [2 Chaps. 7 and 9]). Let n be a positive odd integer, m the multiplicative order of 2 modulo n (we assume that $m \geq 3$), and G the Galois field of order 2^m. Let α be a primitive nth root of unity in G and d a positive integer. The *binary* (*narrow sense*) BCH *code of length n and designed distance d* is the set of all those elements of the quotient algebra $F[x]/(x^n + 1)$, which admit $\alpha, \ldots, \alpha^{d-1}$ as roots. We say that the BCH code of designed distance d is a "true" BCH code if it contains an element that does not admit α^d as root (i.e., if it is different from the BCH code of designed distance $d + 1$).

If n is equal to $2^m - 1$, then the BCH code is called primitive.

A well-known result in algebraic coding theory is the BCH *bound*, which we briefly recount as follows: Let $B(x)$ be an element of $F[x]/(x^n + 1)$ and $f(x)$ its Fourier transform (also called Mattson–Solomon polynomial)

$$f(x) = \sum_{i=1}^{n} B(\alpha^i)x^{n-i}.$$

We have the inverse formula

$$B(x) = \sum_{i=0}^{n-1} f(\alpha^i)x^i.$$

So, $f(x)$ has values 0 and 1 on the set of nth roots of unity and also on zero, since $f(0)$ is equal to $B(1)$. If $n = 2^m - 1$, then f is Boolean and has an even weight (i.e., has a support of an even size). The weight of $B(x)$ is the number of its nonzero terms. It is equal to the number of those nth roots of unity whose images by f are equal to 1. $B(x)$ belongs to the binary BCH code of designed distance d if and only if $f(x)$ has no term whose exponent is equal to $n - d + 1, \ldots, n - 1$, that is, if and only if $d°(f(x)) \leq n - d$. Then $f(x)$ admits at most $n - d$ roots, and the weight of $B(x)$ is at least d. This bound is called the BCH *bound*.

A difficult problem is the characterization of the elements of weight d of this BCH code. Note that the existence of such an element implies that the BCH code is a "true" BCH code. The converse is false, in general.

This problem is related to a problem on the divisibility of polynomials: Let us associate with any element $B(x) = \sum_{i \in I} x^i$ (where I is a subset of $\{0, \ldots, n - 1\}$) of the algebra $F[x]/(x^n + 1)$ the following polynomial over G:

$$\prod_{i \in I} (1 + \alpha^i x^i).$$

This polynomial is called the *locator polynomial* of $B(x)$. We have the following result.

PROPOSITION 1 (see [2, p. 260]). *Any polynomial* $\sigma(x) = \sum_{i=0}^{w} \sigma_i x^i$ *over G is the locator polynomial of a codeword of the* BCH *code of designed distance d if and only if*

(1) $\sigma(0) = 1$,

(2) $\sigma(x)$ *divides* $x^n + 1$,

(3) ($i \in [1, d - 1]$, *i odd*) *implies* ($\sigma_i = 0$).

So $\sigma(x)$ is the locator polynomial of a codeword of weight d of the BCH code of designed distance d if and only if it satisfies conditions (1)–(3) and has degree d.

We are interested in the case where $d = 2^{m-2} - 1$. The main result in [1] says that, if σ is the locator polynomial of an element of weight $2^{m-2} - 1$ of the BCH code of designed distance $2^{m-2} - 1$, then the polynomial $x^{2^{m-2}}\sigma(x^{-1})$ is linearized. So we first rephrase the conditions of Proposition 1 by means of this polynomial.

COROLLARY 1. *Any polynomial $\sigma(x) = \sum_{i=0}^{w} \sigma_i x^i$ is the locator polynomial of a codeword of weight $2^{m-2} - 1$ of the* BCH *code of designed distance $2^{m-2} - 1$ if and only if $x^{2^{m-2}}\sigma(x^{-1})$ is a polynomial that*

(1) *is monic and has degree 2^{m-2},*

(2) *divides $x^{2^m} + x$,*

(3) *has constant nonzero derivative,*

(4) *has zero as a constant term.*

Proof. $\sigma(x)$ has degree $2^{m-2} - 1$ if and only if $x^{2^{m-2}}\sigma(x^{-1})$ is a polynomial that has constant term zero and in which the coefficient of x is nonzero. $\sigma(0)$ is equal to 1 if and only if $x^{2^{m-2}}\sigma(x^{-1})$ is monic and has degree 2^{m-2}. This condition being satisfied, $\sigma(x)$ divides $x^n + 1$ if and only if $x^{2^{m-2}}\sigma(x^{-1})$ divides $x^{2^m} + x$. Condition (3) of Proposition 1, "($i \in [1, 2^{m-2} - 2]$, i odd) implies ($\sigma_i = 0$)," is equivalent to the fact that $x^{2^{m-2}}\sigma(x^{-1})$ has constant derivative. □

We determine explicitly these polynomials in §4. The next section is devoted to the introduction of the necessary mathematical tools.

3. The valuation-representation of polynomials, the separableness. A usual way to express polynomials over a finite field GF(q) is the component representation (cf. [3, p. 38])

$$\sum_{j=0}^{p-1} x^j (P_j(x))^p,$$

where p is the characteristic of the field.

Here, the characteristic being 2, this representation reduces to $(P_0(x))^2 + x(P_1(x))^2$. It results only in a separation between the odd and even exponents in the polynomial. We use a more precise version based on the 2-valuations of the exponents (recall that the 2-valuation of an integer k is the greatest integer j such that 2^j divides k). This representation is obtained by gathering in a same polynomial (for any integer j) all those terms whose exponents are divisible by 2^j and not by 2^{j+1} (j ranging over $\mathbf{N}$). For any j, the corresponding polynomial is equal to the product of x^{2^j} with a polynomial whose exponents are divisible by 2^{j+1}, so it has the form $(xP_j^2(x))^{2^j}$.

DEFINITION 1. *We call valuation-representation of a polynomial $P(x)$ over G the expansion $P(x) = P(0) + \sum_{j\geq 0} (xP_j^2(x))^{2^j}$, where $P_0, \ldots, P_j, \ldots$ are polynomials over G.*

Example. Let $P(x)$ be the following polynomial over G: $\sum_{i=0}^{11} x^i$ (for convenience, all its coefficients are equal to 1). Then $P_0(x)$ is equal to $1 + x + x^2 + x^3 + x^4 + x^5$; $\mathrm{P}_1(x)$ is equal to $1 + x + x^2$; $P_2(x)$ is equal to 1; and $P_3(x)$ is equal to 1. So

$$1 + xP_0^2(x) + (xP_1^2(x))^2 + (xP_2^2(x))^4 + (xP_3^2(x))^8 = P(x).$$

Let us recall that the binary expansion of any positive integer n is

$$n = \sum_{i=0}^{k} \varepsilon_i 2^i, \text{ where } k \text{ is an integer and the coefficients } \varepsilon_i \text{ are equal to 0 or 1.}$$

It is written $\overline{\varepsilon_k \varepsilon_{k-1} \cdots \varepsilon_1 \varepsilon_0}$.

DEFINITION 2. *A polynomial over G is called separable if the exponents of all its nonzero terms admit a binary expansion where the first "1" (at the right hand) is separated from the others (if they exist) by some zeros (at least one).*

Examples. x^3 is not separable, since $3 = \overline{11}$; x^5 is separable, since $5 = \overline{101}$.

More generally, the monomials $x^{2^i}(i \in \mathrm{N})$, $x^{2^i+2^j}(i, j \in \mathrm{N}; j > i + 1)$ are separable, and the monomials $x^{2^i+2^{i+1}}(i \in \mathrm{N})$ are not separable.

Using the valuation-representation, we have the following result.

PROPOSITION 2. *A polynomial over* G: $P(x) = P(0) + \sum_{j \geq 0} (xP_j^2(x))^{2^j}$ *is separable if and only if all the polynomials* $P_j(x)$ *are even (i.e., admit even exponents only).*

The following property will be useful when studying the divisibility of polynomials.

PROPOSITION 3. *Let* $Q(x) = S(x)R(x)$, *where* $S(x)$ *and* $Q(x)$ *are separable polynomials, and i a positive integer. Then, if*

(1) $R(x)$ *is the* 2^i*th power of a polynomial (i.e., all the exponents in* $R(x)$ *admit* 2-*valuations at least equal to i*),

(2) $S_{i-1}(x) \neq 0$ (*where* $S_{i-1}(x)$ *is the polynomial of index* $i - 1$ *of the valuation-representation of* $S(x)$),

then $R(x)$ *is the* 2^{i+1}*th power of a polynomial.*

Proof. Let $R(x) = M^{2^i}(x)$ and $S(x) = S(0) + \sum_{j \geq 0} (xS_j^2(x))^{2^j}$, we have

$$S(x)R(x) = \sum_{j=0}^{i-1} (xM^{2^{i-j}}S_j^2(x))^{2^j} + M^{2^i}(x)\left(S(0) + \sum_{j \geq i} (xS_j^2(x))^{2^j}\right).$$

The polynomial $(SR)_{i-1}(x)$ of the valuation-representation of $S(x)R(x)$ is therefore equal to $S_{i-1}(x)M(x)$.

From $S(x)R(x) = Q(x)$, we deduce that

$$S_{i-1}(x)M(x) = Q_{i-1}(x).$$

$S(x)$ and $Q(x)$ being separable, $S_{i-1}(x)$ and $Q_{i-1}(x)$ are even, and, $S_{i-1}(x)$ being different from 0, $M(x)$ is then even. So $R(x)$ is the 2^{i+1}th power of a polynomial. □

COROLLARY 2. *Let* $Q(x) = S(x)R(x)$, *where* $S(x)$ *and* $Q(x)$ *are separable. Then, if*

(1) $R(x)$ *is even,*

(2) *For all* $i \in N^*$, $S_{i-1}(x) = 0 \Rightarrow R_i(x) = 0$,

then $R(x)$ *is a constant.*

Proof. Let us prove by induction that, for any integer i, $R(x)$ is the 2^ith power of a polynomial; that will prove that $R(x)$ is a constant.

Since $R(x)$ is even, the assertion is true when i equals 1.

Suppose it is true for $i \geq 1$, then we have the following:

(i) If $S_{i-1}(x) \neq 0$, then, according to Proposition 3, $R(x)$ is the 2^{i+1}th power of a polynomial;

(ii) Otherwise, according to hypothesis (2) of the above corollary, $R_i(x) = 0$, and the same conclusion holds. □

The derivative $\sum ia_ix^{i-1} = \sum_{i \text{ odd}} a_ix^{i-1}$ of a separable polynomial $\sum a_ix^i$ is not necessarily separable. We now introduce a stronger property, which is respected by the derivative.

DEFINITION 3. *A polynomial* $P(x)$ *possesses property* (P) *if all the exponents in* $P(x)$ *admit a binary expansion* $\sum_{i \geq 0} \varepsilon_i 2^i$, *where, for any* $i \geq 0$, *either* ε_i *or* ε_{i+1} *is zero.*

Clearly, property (P) implies separableness, and differentiation respects it.

PROPOSITION 4. *Let* $f(x) = \mathrm{tr}(g(x)) \bmod (x^{2^m} + x)$ *be the polynomial expression of a Boolean function. If* $d°f(x) < 2^{m-1} + 2^{m-2}$, *then* $f(x)$ *possesses property* (P).

Proof. Let k be the exponent of any nonzero term of $f(x)$. The elements

$$2^ik \bmod (2^m - 1) \qquad (i \geq 1)$$

of the cyclotomic class of k modulo $(2^m - 1)$ are also the exponents of some nonzero terms of $f(x)$. Let $\sum_{i=0}^{m-1} \varepsilon_i 2^i$ be the binary expansion of k. The binary expansions of the elements of the cyclotomic class of k are obtained from that of k by applying the cyclic permutations on the coefficients ε_i. Suppose that two consecutive ε_i are equal to 1. Then the cyclotomic class of k modulo $(2^m - 1)$ would contain elements at least equal to $2^{m-1} + 2^{m-2}$, and $f(x)$ would not have degree smaller than $2^{m-1} + 2^{m-2}$. So $f(x)$ possesses property (P). □

4. The divisors of $x^{2^m} + x$ of constant derivatives and degree 2^{m-2}. If $m = 2$, then $2^{m-2} = 1$. Any polynomial of degree 1 over GF(4) divides $x^4 + x$. We henceforth assume that $m \geq 3$.

Clearly, if a polynomial admits the zero function as derivative (i.e., if it is an even polynomial), then it cannot divide $x^{2^m} + x$: all its roots are multiple.

So any divisor of $x^{2^m} + x$ of constant derivative is equal, up to the multiplication by an appropriate nonzero element of G, to a polynomial of the form $P^2(x) + x$.

We relate the characterization of such polynomials of degree 2^{m-2}, which divide $x^{2^m} + x$ to the resolution of a class of differential equations.

PROPOSITION 5. *Let $P(x)$ be a polynomial of degree 2^{m-3} $(m \geq 3)$. The polynomial $P^2(x) + x$ divides $x^{2^m} + x$ if and only if $P(x)$ is a solution of one of the differential equations*

$$(1) \qquad ay^4 + y^2y' + y + xy' = x^{2^{m-1}} + ax^2, \qquad (a \in G,\ a \neq 0).$$

Proof. If $P^2(x) + x$ divides $x^{2^m} + x$, then it divides $x^{2^m} + x + P^2(x) + x = (x^{2^{m-1}} + P(x))^2$. Since it is fully reducible and admits no multiple factors, it divides $x^{2^{m-1}} + P(x)$. So there exists a polynomial $Q(x)$ of degree 2^{m-2} such that

$$(2) \qquad x^{2^{m-1}} + P(x) = Q(x)(P^2(x) + x).$$

By differentiating that equality, we obtain

$$(3) \qquad P'(x) = Q'(x)(P^2(x) + x) + Q(x).$$

We have $d°(P'(x)) < 2^{m-3}$ and $d°(P^2(x) + x) = d°(Q(x)) = 2^{m-2}$. Therefore, $Q'(x)$ must be a constant different from zero, say $Q'(x) = a \neq 0$. Equality (3) gives $Q(x) = a(P^2(x) + x) + P'(x)$, and, replacing $Q(x)$ by this value in (2), we obtain $x^{2^{m-1}} + P(x) = (aP^2(x) + ax + P'(x))(P^2(x) + x)$. So $P(x)$ satisfies

$$aP^4(x) + P^2(x)P'(x) + P(x) + xP'(x) = x^{2^{m-1}} + ax^2.$$

$P(x)$ is a solution of (1).

Conversely, if $P(x)$ satisfies (1), then $P^2(x) + x$ divides $x^{2^{m-1}} + P(x)$ and therefore divides $x^{2^m} + x$. □

PROPOSITION 6. *Let a be any nonzero element of $G =$ GF(2^m) $(m \geq 3)$. The solutions of the differential equation over G,*

$$(4) \qquad ay^4 + y^2y' + y + xy' = x^{2^{m-1}} + ax^2,$$

are all affine polynomials (i.e., are the sums of linearized polynomials and constants).

Proof. Let $P(x)$ be a solution of (4). If $P'(x) = 0$, then $P(x)$ satisfies the relation $aP^4(x) + P(x) = x^{2^{m-1}} + ax^2$ and is therefore an affine polynomial. We henceforth assume that $P'(x) \neq 0$.

We first prove that $P(x)$ is separable. Set $h(x) = P'(x) + aP^2(x) + ax$. We have

$$\begin{aligned}(P^2(x) + x)h(x) &= (P^2(x) + x)(P'(x) + aP^2(x) + ax)\\ &= aP^4(x) + P^2(x)P'(x) + xP'(x) + ax^2\\ &= P(x) + x^{2^{m-1}}\end{aligned}$$

(according to (4)). Let $g(x) = P(x)h(x)$ and $f(x) = g(x) + x^{2^{m-1}}h(x) = (P(x) + x^{2^{m-1}})h(x)$.

According to the preceding equality, $f(x)$ is equal to $(P^2(x) + x)h^2(x) = g^2(x) + xh^2(x)$, and $f(x) + f^2(x)$ is therefore equal to

$$g^2(x) + xh^2(x) + (g(x) + x^{2^{m-1}}h(x))^2 = (x^{2^m} + x)h^2(x).$$

$f(x) + f^2(x)$ is divisible by $x^{2^m} + x$, so $f(x)$ is the polynomial expression of a Boolean function. Its degree is $2^{m-1} + 2^{m-2}$.

The number $2^{m-1} + 2^{m-2}$ being the greatest element of the cyclotomic class of 3, there exists an element λ of G such that $f(x) = \operatorname{tr}(\lambda x^3) \bmod (x^{2^m} + x) + f_1(x)$, where $f_1(x)$ is the polynomial expression of a Boolean function of degree smaller than $2^{m-1} + 2^{m-2}$ (λ is such that the coefficient of $x^{2^{m-1}+2^{m-2}}$ in $f(x)$ is $\lambda^{2^{m-2}}$). According to Proposition 4, $f_1(x)$ and $f'_1(x) = h^2(x) + \lambda x^2 + (\lambda x)^{2^{m-1}}$ (and therefore $h(x)$) possess property (P). We deduce that $P(x)$ has property (P), since, if it was not the case, then, by considering the greatest exponent in $P(x)$ whose binary expansion has two consecutive 1's, we would deduce that $h(x)$ does not possess (P) and arrive at a contradiction.

So, $P(x)$ is separable. We now prove that $P'(x)$ is a constant. If $m = 3$, then it is obvious. So, we may assume that $m > 3$.

$P(x)$ being a solution of (4), we have

$$(5) \qquad P^2(x)P'(x) = aP^4(x) + P(x) + xP'(x) + x^{2^{m-1}} + ax^2.$$

Let us prove that the polynomials $S(x) = P^2(x)$, $R(x) = P'(x)$, and $Q(x) = aP^4(x) + P(x) + xP'(x) + x^{2^{m-1}} + ax^2$ satisfy the hypothesis of Corollary 2.

- $P(x)$ being separable, $S(x)$ and $Q(x)$ are separable.
- $R(x)$ is even.
- Suppose that, for a positive integer i, the polynomial S_{i-1}, which is equal to P_{i-2}, is the zero polynomial; then we must prove that $R_i = 0$.

Since $d^\circ P(x)$ is 2^{m-3}, we may write $P(x) = \mu x^{2^{m-3}} + T(x)$, where $d^\circ T < 2^{m-3}$. Equality (5) gives $a\mu^4 = 1$ and

$$[\mu^2 x^{2^{m-2}} + T^2(x)]T'(x) = aT^4(x) + T(x) + xT'(x) + ax^2 + \mu x^{2^{m-3}}.$$

Therefore, since $T'(x) = P'(x) \neq 0$, $d^\circ T^4(x)$ is equal to $2^{m-2} + d^\circ T'(x)$ and $4d^\circ T(x) < 2^{m-2} + d^\circ T(x)$; $d^\circ T(x) < 2^{m-2}/3$. Thus $d^\circ T^2(x)T'(x) < 2^{m-2}$. Therefore, since the degree of $T(x) + xT'(x) + ax^2 + \mu x^{2^{m-3}}$ is smaller than 2^{m-2}, the polynomial $\mu^2 x^{2^{m-2}} T'(x)$ is equal to the sum of the terms of degrees at least 2^{m-2} of the polynomial $aT^4(x)$.

If $i \geq m - 2$, then, since $R(x)$ has degree smaller than 2^{m-3}, $R_i(x) = 0$. So we may assume that $i < m - 2$. Since, by hypothesis P_{i-2} is zero, $P(x)$ has no exponent of 2-valuation $i - 2$, $T^4(x)$ has no exponent of 2-valuation i, and so $\mu^2 x^{2^{m-2}}\ T'(x)$ has no exponent of 2-valuation i. Thus $R_i(x) = 0$ (since $i < m - 2$).

We have proved that $P'(x)$ is a constant.

Let $P'(x) = b$. $P(x)$ is then a solution of the polynomial equation

$$(6) \qquad aP^4(x) + bP^2(x) + P(x) = x^{2^{m-1}} + ax^2 + bx$$

and is therefore affine. □

We now give the explicit expressions of those polynomials that satisfy the conditions stated in the title of the paper.

THEOREM. *Let $m \geq 2$ and G the Galois field of order 2^m. Those polynomials over G that divide $x^{2^m} + x$, whose derivatives are constant, and whose degree is 2^{m-2} are all the polynomials of the type*

$$k(u \operatorname{tr}(v^2 x) + v \operatorname{tr}(u^2 x) + \varepsilon u + \eta v),$$

where u and v are two linearly independent elements of the F-space G, k is any nonzero element of G, and ε, $\eta \in \{0, 1\}$.

Proof. Clearly, any polynomial of this form divides $x^{2^m} + x$ since

(i) It has degree 2^{m-2}; indeed, the coefficient of $x^{2^{m-1}}$ is equal to $k(uv + vu) = 0$, and the coefficient of $x^{2^{m-2}}$ is equal to $k(uv^{2^{m-1}} + vu^{2^{m-1}}) = k(uv(u + v))^{2^{m-1}} \neq 0$ (since u and v are linearly independent);

(ii) Its roots are all the elements of the $(m - 2)$-dimensional flat

$$\{x \in G/\operatorname{tr}(v^2x) = \varepsilon \text{ and } \operatorname{tr}(u^2x) = \eta\}.$$

Conversely, if $m = 2$, then it is a simple matter to check that any polynomial of degree 1 over G admits such a representation. So we may restrict ourselves to $m \geq 3$.

Let $M(x)$ be a polynomial of degree 2^{m-2}, which admits a constant derivative and divides $x^{2^m} + x$. We have seen that we may write $M(x) = k(P^2(x) + x)$ $(k \in G, k \neq 0)$.

Propositions 5 and 6 prove that $P(x)$ is affine. So $M(x)$ is affine. We may restrict ourselves to the case where $M(x)$ is a linearized polynomial (if $M(x)$ is an affine polynomial that divides $x^{2^m} + x$, then, for any root a of $M(x)$, the polynomial $M(x + a)$ is linearized, and the set of polynomials of the form $k(u \operatorname{tr}(v^2x) + v \operatorname{tr}(u^2x) + \varepsilon u + \eta v)$ is invariant under any translation).

$M(x)$ is now a linearized polynomial whose roots are the elements of an $(m - 2)$-dimensional subspace E of G. There exists u and v, linearly independent, such that E is equal to $\{x \in G/\operatorname{tr}(v^2x) = 0 \text{ and } \operatorname{tr}(u^2x) = 0\}$. The polynomial $k(u \operatorname{tr}(v^2x) + v \operatorname{tr}(u^2x))$ admits E as a set of roots and has degree 2^{m-2}. Thus, by unicity of the (monic) linearized polynomial that admits E as a set of roots, $M(x)$ is equal to $k(u \operatorname{tr}(v^2x) + v \operatorname{tr}(u^2x))$. □

As stated in the Introduction, we deduce a result equivalent to that of Augot, Charpin, and Sendrier [1].

COROLLARY 3. *Let $m \geq 3$ and $n = 2^m - 1$. Let $G = \mathrm{GF}(2^m)$. Let $B(x)$ be any element of the algebra $F[x]/(x^n + 1)$. Then $B(x)$ belongs to the* BCH *code of length n and designed distance $2^{m-2} - 1$ and has weight $2^{m-2} - 1$ if and only if its locator polynomial $\sigma(x)$ satisfies*

$$x^{2^{m-2}}\sigma(x^{-1}) = \frac{1}{uv^{2^{m-1}} + vu^{2^{m-1}}}(u \operatorname{tr}(v^2x) + v \operatorname{tr}(u^2x)),$$

where u and v are two linearly independent elements of G.

REFERENCES

[1] D. AUGOT, P. CHARPIN, AND N. SENDRIER, *Sur une Classe de Polynômes Scindés de l'Algèbre $F_{2^m}[Z]$*, C.R. Acad. Sci. Paris, t.312, Série I, 1991, pp. 649–651; *Studying the locator polynomials of minimum weight codewords of* BCH *codes*, IEEE Trans. Inform. Theory, 38 (1992), pp. 960–973.

[2] F. J. MAC WILLIAMS AND N. J. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[3] L. REDEI, *Lacunary Polynomials over Finite Fields*, Akadémiai Kiado, Budapest, Hungary, 1973.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 245–259, May 1994

THE k-EDGE-CONNECTED SPANNING SUBGRAPH POLYHEDRON*

SUNIL CHOPRA†

Abstract. This paper studies the polyhedron $P_k(G)$ defined by the convex hull of k-edge-connected spanning subgraphs of a given graph G where multiple copies of an edge are allowed. A complete inequality description of $P_k(G)$ when k is odd and G is an outer planar graph is given. A family of facet-defining inequalities of $P_k(G)$ that have the same support graph but coefficients that depend on $k \in \{4r - 2, 4r - 1, 4r + 1, r \in \{1, 2, \dots\}\}$ is described.

Key words. k-edge-connected, polyhedron, facet, outerplanar graph

AMS subject classifications. 05C40, 90C27

1. Introduction. In the design of communication or transportation networks, it is frequently important to produce networks of low "cost" that are "survivable." Several definitions of survivability are discussed in Grötschel and Monma [10]. Stoer [14] studied these problems for the case when no more than one copy of an edge may be used. She describes various decompositions for sparse input graphs. She also gives several valid and facet-defining inequalities for the associated polyhedra. The problem where multiple edges are not allowed was also studied by Christofides and Whitlock [4], Monma, Munson, and Pulleyblank [12], and Bienstock, Brickell, and Monma [1]. In this paper, we only examine the case where the survivability requirements are given in terms of edge-connectivity and where multiple copies of an edge are allowed. If edge-connectivity is the survivability criterion used, allowing multiple copies of an edge often results in a lower-cost solution than the case when multiple copies are not allowed. Thus the study of this problem is important. The instance where multiple copies of an edge are allowed is a relaxation of the case in which at most one copy is used. Thus all valid inequalities identified by us are also valid for the case when at most one copy of an edge may be used.

A *graph* $G = (V, E)$ is called *k-edge-connected* if the removal of any $(k - 1)$ or fewer edges leaves G connected. Note that we allow multiple copies of a given edge. A *family of edges* of E is a collection F of elements of E. Several copies of the same edge may appear in F. For each edge $e \in F$, the multiplicity of e in F is the number of times e appears in F. An element of a vector x indexed by the edge set E is referred to as x_e or $x(e)$. With every family F of edges of G, we associate a unique incidence vector $x^F \in R^E$ by setting x_e^F equal to the multiplicity of e in F. Clearly, if $e \notin F$, then $x_e^F = 0$. By $G[F]$, we denote the multigraph obtained by replicating each edge $e \in F$ as many times as its multiplicity in F. Given edge weights c_e, $e \in E$, the *weight* of a family F is given by $c(F)$, where

$$c(F) = \sum_{e \in E} c_e x_e^F.$$

In this paper, we consider the *k-edge-connected spanning subgraph problem* (k-ECSSP): Given a graph $G = (V, E)$ with nonnegative edge weights $c_e \in R^+$, find a family F of edges of minimum weight such that $G[F]$ spans G and is k-edge-connected. For $k \geq 2$,

* Received by the editors November 25, 1991; accepted for publication (in revised form) April 28, 1993.

† Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208.

this problem is NP-hard (see Garey and Johnson [7]). For $k = 1$, the problem reduces to that of finding a minimum-weight spanning tree that can be solved in polynomial time.

Given a connected graph $G = (V, E)$, define the *k-edge-connected spanning subgraph polyhedron* $P_k(G)$, where

$$P_k(G) = \text{conv}\ \{x^F \mid G[F] \text{ spans } G \text{ and is } k\text{-edge-connected}\}.$$

Note that $P_k(G)$ is a polyhedron and contains the nonnegative orthant as its recession cone. k-ECSSP can be solved as min $\{c^T x \mid x \in P_k(G)\}$. We study the polyhedron $P_k(G)$ for $k \geq 2$. In §2 we give an integer linear programming formulation of the problem. Cornuéjols, Fonlupt, and Naddef [5] showed that the corresponding LP-relaxation defines $P_k(G)$ for k even and G a series-parallel graph. We show that, for k odd, the LP-relaxation has fractional extreme points even if G is a triangle. For k odd, we introduce the class of *outerplanar partition inequalities* and show them to be facet-defining for $P_k(G)$. In §3 we show that, for k odd, the LP-relaxation, along with the outerplanar partition inequalities, defines $P_k(G)$ if G is outerplanar. In §4 we introduce other facet-defining inequalities for $P_k(G)$. Finally, we show how facet-defining inequalities for $P_k(G)$ can be used to obtain facet-defining inequalities for high edge-connectivity network design problems. For basic definitions and results in polyhedral combinatorics, refer to Pulleyblank [13].

2. Formulation and outerplanar partition inequalities. For each edge $e \in E$, we introduce a variable x_e, referring to the multiplicity of e in the solution. An edge e with end nodes i and j may also be referred to as (i, j). Given $\bar{V} \subseteq V$, define the cut

$$\delta(\bar{V}) = \{e \in E \mid e = (i, j) \text{ where } i \in \bar{V}, j \in V - \bar{V}\}.$$

For $1 \leq |\bar{V}| \leq |V| - 1$, the *cut-inequality*

$$\sum_{e \in \delta(\bar{V})} x_e \geq k \tag{2.1}$$

is valid for $P_k(G)$. k-ECSSP can be formulated as the following integer linear program:

$$\min \{c^T x \mid x \text{ satisfies (2.1) for all } \bar{V} \subseteq V,\ 1 \leq |\bar{V}| \leq |V| - 1,\ x \geq 0, \text{ and integral}\}.$$

Define the polyhedron $LP_k(G)$, where

$$LP_k(G) = \{x \in R_+^E \mid x \text{ satisfies (2.1) for all } \bar{V} \subseteq V,\ 1 \leq |\bar{V}| \leq |V| - 1\}.$$

Since $P_k(G)$ contains the nonnegative orthant as its recession cone, it is clearly full-dimensional. Using standard polyhedral techniques, we can show the following result.

PROPOSITION 2.1. *Let $G = (V, E)$ be any connected graph. Then* (i) *the cut-inequality* (2.1) *is facet-defining for $P_k(G)$ if and only if both $\bar{V}$ and $V - \bar{V}$ induce connected subgraphs of G*; (ii) *the inequality $x_e \geq 0$ is facet-defining for $P_k(G)$ if and only if there does not exist $\bar{V} \subseteq V$ such that $\delta(\bar{V}) = \{e\}$.* □

This formulation was considered by Goemans [8] and Goemans and Bertsimas [9]. They gave a worst-case bound of the ratio between the optimal solution to the LP-relaxation and the integer optimum. Goemans in [8] also provided a probabilistic analysis of this ratio in the case where the nodes of the graph are uniformly distributed in the unit cube. This ratio was shown to have a limit that can be bounded away from one. This emphasizes the need for additional inequalities to strengthen the formulation.

Another way to check the strength of the cut-inequalities is to characterize the class of graphs for which $P_k(G) = LP_k(G)$. Cornuéjols, Fonlupt, and Naddef [5] showed the following result.

THEOREM 2.1. *For k even, if G is a series-parallel graph, then $P_k(G) = LP_k(G)$ and $x_e \in \{0, k/2, k\}$ for all $e \in E$, where x is a vertex of $P_k(G)$.*

The above result does not hold for k odd. As an example, consider a triangle induced by edges e_1, e_2, e_3. The solution $x_e = k/2$ for $e \in \{e_1, e_2, e_3\}$ is a fractional vertex of $LP_k(G)$. However, the following result follows from Theorem 2.1.

PROPOSITION 2.2. *If x is a vertex of $LP_k(G)$ for k odd and G is a series-parallel graph, then $x_e \in \{0, k/2, k\}$ for all $e \in E$.*

This may lead us to believe that, for G series-parallel, $P_k(G)$ has extreme points that can only take on a limited number (not all integers between 0 and k) of integer values for k odd. This proves to be false. Given any integer $0 \leq r \leq k$, we can construct a series-parallel graph for which $P_k(G)$ has an extreme point x with $x_e = r$ for some edge e. For outerplanar graphs, we can show that, if x is an extreme point of $P_k(G)$, then $x_e \in \{0, 1, \lfloor k/2 \rfloor, \lceil k/2 \rceil, k - 1, k\}$ for all $e \in E$.

For the remainder of this section and the next section, we assume that k is odd and $k \geq 3$. We define two basic operations of contraction and deletion. Given a graph $G = (V, E)$ and an edge $e = (u, v)$, *contracting* the edge e consists of identifying the nodes u and v into the node u and deleting edge e to get $G_c = (V_c, E_c)$. All edges that were incident to either u or v in G are now incident to the combined node u. The reverse operation is called *expansion. Deleting* edge e results in the graph $G_d = (V_d, E_d)$, where $V_d = V$ and $E_d = E - \{e\}$. The reverse operation is called *extension.* We obtain a *minor* G' of G by a sequence of contractions and deletions. We give basic lifting theorems associated with expansion and extension. They are stated here without proof. Let $G_c = (V_c, E_c)$ be the graph obtained from $G = (V, E)$ by contracting edge $\bar{e}$.

THEOREM 2.2. *If the inequality $\sum_{e \in E - \{\bar{e}\}} a_e x_e \geq a_0$ is facet-defining (valid) for $P_k(G_c)$, then $\sum_{e \in E} a_e x_e \geq a_0$ is facet-defining (valid) for $P_k(G)$, where $a(\bar{e}) = 0$.*

In the case of extension, lifting is more complex (can be as difficult to solve as the original problem). However, there is one interesting instance where lifting is easy. Let $G = (V, E)$ be a graph with two parallel edges e_1 and e_2. G_d is obtained from G by deleting e_2.

THEOREM 2.3. *If the inequality $\sum_{e \in E - \{e_2\}} a_e x_e \geq a_0$ is facet-defining (valid) for $P_k(G_d)$, then $\sum_{e \in E} a_e x_e \geq a_0$ is facet-defining (valid) for $P_k(G)$, where $a(e_2) = a(e_1)$.*

The above results can be used to prove the following more general lifting theorem. Consider any partition $\pi = (V_i, i = 1, \ldots, r)$ of the node set V. Assume that each of the subsets V_i induces a connected subgraph $G_i = (V_i, E_i)$ of G. Let $G_\pi = (V_\pi, E_\pi)$ be the graph obtained by shrinking each subset V_i into a single node v_i. Let T_i be a spanning tree of the graph G_i, $i = 1, \ldots, r$. G_π is obtained from G by contracting the edges in T_i, $i = 1, \ldots, r$ and deleting the edges $E_i - T_i$, $i = 1, \ldots, r$. G_π is referred to as a *contraction minor* of G. Repeated applications of Theorems 2.2 and 2.3 prove the following result.

THEOREM 2.4. *If the inequality $\sum_{e \in E_\pi} a_e x_e \geq a_0$ is facet-defining (valid) for $P_k(G_\pi)$, then $\sum_{e \in E} a_e x_e \geq a_0$ is facet-defining (valid) for $P_k(G)$, where $a_e = 0$ for $e \in E - E_\pi$.*

We now turn to the fractional extreme point of $P_k(G)$, where G is a triangle. This extreme point is cut off by the facet-defining inequality

$$x(e_1) + x(e_2) + x(e_3) \geq 3\lceil k/2 \rceil - 1.$$

We generalize this inequality to outerplanar graphs. A graph $G = (V, E)$ is said to be outerplanar if it is planar and can be embedded on the plane so that all nodes lie on the outermost face. A well-known property of outerplanar graphs is as follows (see, for instance, Gan and Johnson [6]).

PROPOSITION 2.3. *Any connected outerplanar graph $G = (V, E)$ with $|V| \geq 2$ contains one of the configurations in Fig. 2.1.*

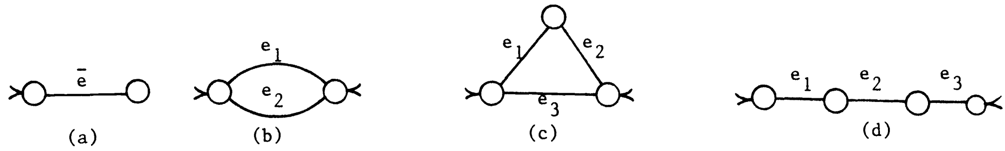


FIG. 2.1

Let $G = (V, E)$ be any outerplanar graph. The *outerplanar partition inequality* (OP-inequality) is given by

$$(2.2) \qquad \sum_{e \in E} x_e \geq |V| \lceil k/2 \rceil - 1.$$

LEMMA 2.1. *The* OP-*inequality* (2.2) *is valid for* $P_k(G)$.

Proof. The proof is by induction on G (i.e., $|E|\,|V|$). Inequality (2.2) is clearly valid if G is an edge. Assume that inequality (2.2) is valid for all minors G' of G. If G contains Fig. 2.1(a), then $x(\bar{e}) \geq k$ for all $x \in P_k(G)$. Contract the edge $\bar{e}$ to obtain $G' = (V', E')$. Let x' be the restriction of x to E'. By induction,

$$\sum_{e \in E - \{\bar{e}\}} x_e \geq (|V| - 1)\lceil k/2 \rceil - 1.$$

Since $x_e \geq k$, it follows that x satisfies (2.2). If Fig. 2.1(a) is not present, then, by Proposition 2.3, G either contains parallel edges e_1 and e_2 or a node u with degree $(u) = 2$. If G contains parallel edges e_1 and e_2, delete e_2 to obtain G'. By the induction hypothesis,

$$\sum_{e \in E - \{e_2\}} x_e \geq |V| \lceil k/2 \rceil - 1.$$

Since $x(e_2) \geq 0$, the validity of (2.2) follows.

If G contains a node u of degree 2, let e_1 and e_2 be the two edges incident to u. Contract e_1 to obtain G'_1, and e_2 to obtain G'_2. Each defines an OP-inequality. By the induction hypothesis, we have

$$(2.3) \qquad \sum_{e \in E - \{e_1\}} x_e \geq (|V| - 1)\lceil k/2 \rceil - 1.$$

$$(2.4) \qquad \sum_{e \in E - \{e_2\}} x_e \geq (|V| - 1)\lceil k/2 \rceil - 1.$$

From the cut-inequalities, we have

$$(2.5) \qquad x(e_1) + x(e_2) \geq k.$$

Adding (2.3)–(2.5) and dividing by 2 gives

$$(2.6) \qquad \sum_{e \in E} x_e \geq (|V| - 1)\lceil k/2 \rceil + (k/2) - 1.$$

On rounding the right side of (2.6), we obtain (2.2). This proves the result. □

Define a graph G to be 2-*connected* if the removal of any node and all incident edges leaves G connected.

THEOREM 2.5. *The* OP-*inequality* (2.2) *is facet-defining for* $P_k(G)$ *if and only if* G *is* 2-*connected.*

Proof. If $G = (V, E)$ is not 2-connected, there is a node v that is a cut-node for G, as shown in Fig. 2.2.

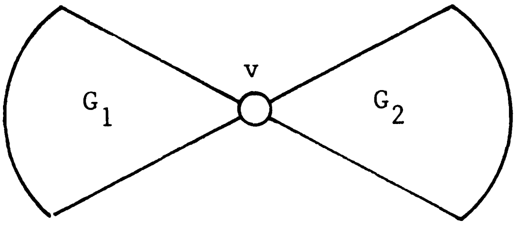


FIG. 2.2

Assume that $G_i = (V_i, E_i)$, $i = 1, 2$. Each of G_1 and G_2 is outerplanar and gives rise to OP-inequalities

$$\sum_{e \in E_i} x_e \geq |V_i| \lceil k/2 \rceil - 1 \quad \text{for } i = 1, 2$$

that are valid for $P_k(G)$ by Theorem 2.4. Since $|V| = |V_1| + |V_2| - 1$, adding the OP-inequalities defined by G_1 and G_2 gives

$$\sum_{e \in E} x_e \geq (|V_1| + |V_2|)\lceil k/2 \rceil - 2 = (|V| + 1)\lceil k/2 \rceil - 2,$$

which has a larger right side than (2.2), since $k \geq 3$. Thus (2.2) cannot be facet-defining in this case.

Now consider the case where G is 2-connected. The nodes in V can be ordered from 1 to $|V|$ such that the nodes i, $i + 1$ are adjacent on the outermost face of G. Define the edge $e_i = (i, i + 1)$ for $i = 1, \ldots, |V| - 1$, with $e_{|V|} = (|V|, 1)$ (if there are several parallel edges $(i, i + 1)$ any one may be considered to be e_i). Consider any valid inequality $\alpha x \geq \alpha_0$ such that

$$\{x \in P_k(G) \mid x \text{ satisfies (2.2) with equality}\} \subseteq \{x \mid \alpha x = \alpha_0\}.$$

Given $1 \leq i \leq |V|$, consider the vector x^i, where

$$x_e^i = \begin{cases} \lfloor k/2 \rfloor & \text{for } e = e_i, \\ \lceil k/2 \rceil & \text{for } e = e_j, \quad j \in \{1, \ldots, |V|\} - \{i\}, \\ 0 & \text{otherwise.} \end{cases}$$

$x^i \in P_k(G)$ and satisfies (2.2) with equality for $i \in \{1, \ldots, |V|\}$. This shows that

$$\alpha(e_i) = \alpha(e_j) = b \quad \text{for } i, j \in \{1, \ldots, |V|\}, \quad i \neq j.$$

Now consider an edge $\bar{e} = (r, s) \in E$, where $s \notin \{r - 1, r + 1\}$, indices calculated modulo $|V|$. Define the vector $\bar{x}$, where

$$\bar{x}_e = \begin{cases} \lfloor k/2 \rfloor & \text{for } e = e_s, \\ x_e^r & \text{for } e \in E - \{e_s, \bar{e}\}, \\ 1 & \text{for } e = \bar{e}. \end{cases}$$

$\bar{x} \in P_k(G)$ and satisfies (2.2) with equality. Comparing $\bar{x}$ and x^r, we have

$$\alpha(\bar{e}) = \alpha(e_s) = b.$$

Since $\bar{e}$ is an arbitrary edge from $E - \{e^i, i \in \{1, \ldots, |V|\}\}$, this shows that $\alpha_e = b$ for all $e \in E$. Thus $\alpha x \geq \alpha_0$ is a multiple of (2.2), and the result follows by standard polyhedral theory. □

We now describe a procedure for lifting OP-inequalities. This procedure was suggested by Stoer. Let $G = (V, E)$ be any outerplanar graph and let $\bar{G} = (V, \bar{E})$ be the graph obtained on adding edges $\bar{E} - E = \{e_1, \dots, e_l\}$. Let inequality (2.2) be the OP-inequality that is facet-defining for $P_k(G)$. The *lifted outerplanar partition* inequality (LOP-inequality) for $P_k(\bar{G})$ is given by

$$(2.7) \qquad \sum_{e \in E} x_e + \sum_{j=1}^{l} a(e_j)x(e_j) \geq |V|\lceil k/2 \rceil - 1,$$

where $a(e_j)$ is the number of edges in the shortest path in G linking the end nodes of e_j. The LOP-inequality (2.7) is clearly valid for $P_k(\bar{G})$. However, it is not, in general, facet-defining.

A necessary condition for (2.7) to be facet-defining was provided by Stoer. It is described in the following. First, we need a definition and a result on outerplanar graphs. G is a *maximal outerplanar graph* (MOP-graph) if the addition of any nonparallel edge would make the graph not outerplanar. Let $G = (V, E)$ be a MOP-graph and let $\{u, v\} \subseteq V$ be two nodes such that (u, v) is not an edge in G. The nodes u and v divide the outer face of G into two paths $P_i = (V_i, E_i)$, linking u to v. Here $V_1 \cap V_2 = \{u, v\}$ and P_1 (P_2) is a path from u to v.

LEMMA 2.2. *If G, u, v, P_1, and P_2 are as defined above and the shortest path from u to v in G contains $l \geq 2$ edges, then there exist $l - 1$ edges (u_i, v_i), $i \in \{1, \dots, l-1\}$ such that*

(a) *$u_i \in V_1 - \{u, v\}$ and $v_i \in V_2 - \{u, v\}$ for $i \in \{1, \dots, l-1\}$,*

(b) *$u_i(v_i) \neq u_j(v_j)$ for $i \neq j$.*

Proof. The proof is by induction on l. Consider the case where $l = 2$. Let $w \in V_2$ be the intermediate node on this path with two edges. The path P_1 must have at least two edges. Since $(u, v) \notin E$ and G is a MOP-graph, (w, z) must be an edge for some $z \in V_2 - \{u, v\}$. Thus the result holds in this case.

Assume that the result holds for $l = 2, \dots, r - 1$ for $r \geq 3$. Consider the case where $l = r$. Let $P_s = (V_s, E_s)$ be the shortest path in G from u to v. Note that $|E_s| = r$. Without loss of generality, assume that (w, v) is the last edge in P_s and $w \in V_2$. $\bar{P}_s = (V_s - \{v\}, E_s - \{(w, v)\})$ is the shortest path in G from u to w and has length $r - 1 \geq 2$. By the induction hypotheses, there exist $r - 2$ edges (u_i, v_i), $i = 1, \dots, r-2$ such that

$$u_i \in V_1 - \{u\} \quad \text{and} \quad v_i \in V_2 - \{u, V_2[w, v]\}; \quad u_i(v_i) \neq u_j(v_j) \quad \text{for } i \neq j.$$

Here $V_2[w, v]$ are the set of nodes in P_2 in the segment between w and v (including v and w).

First, we show that $u_{r-2} \neq v$. To the contrary, assume that $v = u_{r-2}$. Thus (v, v_{r-2}) is an edge in G. Since G is an outerplanar graph, each path in G from u to w must pass through either v or v_{r-2}. The shortest path $\bar{P}_s$ does not pass through v. Thus it must pass through v_{r-2}. Then, however, G has a path from u to v of length less than r if we follow $\bar{P}_s$ to v_{r-2} and then use the edge (v, v_{r-2}). This contradiction shows that $u_{r-2} \neq v$. The nodes u, v, w (u_i, v_i, $i = 1, \dots, r-2$) are as shown in Fig. 2.3. The nodes in V_1 have a natural ordering as they appear in P_1. Given $\{s, t\} \subseteq V_1$, let $V_1[s, t]$ correspond to the nodes in V_1 on the segment in P_1 from s to t (including s and t). Since $(v, v_{r-2}) \notin E$ and G is a MOP-graph, there exists an edge (w, s), where $s \in V_1[u_{r-2}, v] - \{v\}$. If $s \notin u_{r-2}$, then (w, s) and the edges (u_i, v_i), $i \in \{1, \dots, r-2\}$ give the $r - 1$ edges required. Thus we need only consider the case where $(w, u_{r-2}) \in E$ and where w is not linked to any node in $V_1[u_{r-2}, v] - \{u_{r-2}, v\}$ by an edge in E. Since G is a MOP-graph, $(v, u_{r-2}) \in E$.

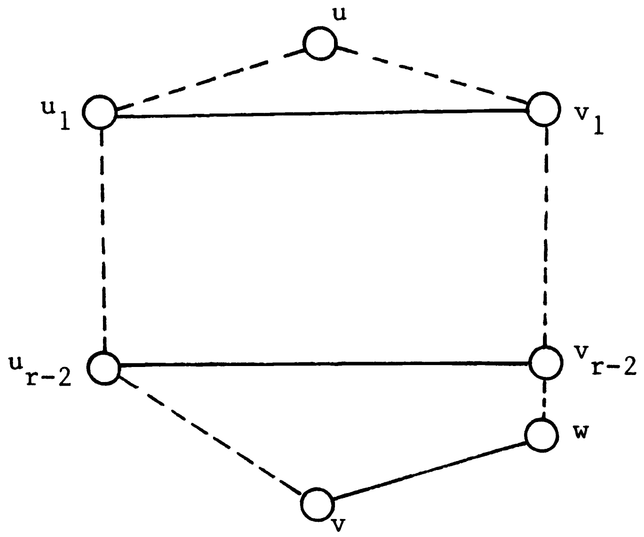


FIG. 2.3

Using the same argument with v_{r-2} (instead of w) shows that we either have the required $r-1$ edges or $\{(v_{r-2}, u_{r-3}), (u_{r-2}, u_{r-3})\} \subseteq E$. Continuing in this manner, we are able to show that either we have the required $r-1$ edges or $\{(v_{r-j}, v_{r-j-1}), (u_{r-j}, u_{r-j-1})\} \subseteq E$ for $j \in \{2, \ldots, r-2\}$. Thus there is a path from u_1 to v of length $r-2$. This implies that $(u, u_1) \notin E$, and there exists at least one node in $V_1[u, u_1] - \{u, u_1\}$. Since G is a MOP-graph and $(u, u_1) \notin E$, there exists an edge $(v_1, t) \in E$ for $t \in V_1[u, u_1] - \{u, u_1\}$. Then the edges (v_1, t), $\{(u_i, v_{i+1}), i = 1, 2, \ldots, r-3\}$, (u_{r-2}, w) give the required $r-1$ edges. The result thus follows. □

We are now able to prove the following result suggested by Stoer.

THEOREM 2.6. *Let $G = (V, E)$ be a* MOP-*graph and let $\bar{G} = (V, \bar{E})$ contain G as a subgraph with $\bar{E} - E = \{e_1, \ldots, e_l\}$. Then the* LOP-*inequality* (2.7) *is facet-defining for $P_k(\bar{G})$.*

Proof. Consider any valid inequality $\alpha x \geq \alpha_0$ such that

$$\{x \in P_k(\bar{G}) \mid x \text{ satisfies (2.7) with equality}\} \subseteq \{x \mid \alpha x = \alpha_0\}.$$

Using the same argument as in the proof of Theorem 2.5, we can show that

$$(2.8) \qquad \alpha_e = b \quad \forall e \in E.$$

Consider any edge $\bar{e} \in \bar{E} - E$, where $\bar{e} = (u, v)$. Let $r \geq 2$ be the length of the shortest path from u to v in G. By Lemma 2.3, we have edges $e_i = (u_i, v_i) \in E$ for $i = 1, \ldots, r-1$ satisfying conditions (a) and (b) in Lemma 2.3. The paths P_1 and P_2 are as defined in Lemma 2.3. For $\{s, t\} \subseteq V_i$, $i = 1, 2$, $P_i[s, t] = (V_i[s, t], E_i[s, t])$ corresponds to the segment of path P_i between s and t ($V_i[s, t]$ includes both s and t). Let $u = u_0 = v_0$ and $v = u_r = v_r$ for ease of notation. Define the solution x, where

$$x_e = \begin{cases} 1 & \text{for } e = e_i, \quad i = 1, \ldots, r-1, \\ \lfloor k/2 \rfloor & \text{for exactly one edge in } E_2[v_i, v_{i+1}], \quad \text{for } i \in \{0, \ldots, r-1\}, \\ \lceil k/2 \rceil & \text{for all other edges in } E_1 \cup E_2, \\ 0 & \text{otherwise.} \end{cases}$$

x satisfies (2.7) with equality. Define $\bar{x}$, where

$$\bar{x}_e = \begin{cases} 1 & \text{for } e = \bar{e}, \\ \lfloor k/2 \rfloor & \text{for exactly one edge } \tilde{e}_i \text{ in } E_1[u_i, u_{i+1}], \quad \text{for } i \in \{0, \dots, r-1\}, \\ x_e & \text{otherwise.} \end{cases}$$

$\bar{x}$ also satisfies (2.7) with equality. Comparing x and $\bar{x}$, we have

$$\alpha(\bar{e}) = \sum_{i=0}^{r-1} \alpha(\tilde{e}_i) = rb.$$

Thus $\alpha x \geq \alpha_0$ is a multiple of (2.7), and the result follows. □

As an example, let $\bar{G} = (V, \bar{E})$ be the complete graph on four nodes with $\bar{E} = \{e_i, i = 1, \dots, 6\}$. Let $G = (V, E)$, where $\bar{E} - E = \{e_6\}$. G is an outerplanar graph, and the OP-inequality

$$\sum_{i=1}^{5} x(e_i) \geq 4\lceil k/2 \rceil - 1$$

is facet-defining for $P_k(G)$. On lifting, we obtain the LOP-inequality

$$\sum_{i=1}^{5} x(e_i) + 2x(e_6) \geq 4\lceil k/2 \rceil - 1$$

that is facet-defining for $P_k(\bar{G})$.

3. The polyhedron $P_k(G)$ of an outerplanar graph for k odd. Let $G = (V, E)$ be any outerplanar graph. Each partition $\pi = (V_i, i = 1, \dots, r)$, where V_i induces a connected subgraph of G for $i = 1, \dots, r$, defines a contraction minor $G_\pi = (V_\pi, E_\pi)$ of G. G_π is outerplanar and defines an OP-inequality

$$\sum_{e \in E_\pi} x_e \geq |V_\pi| \lceil k/2 \rceil - 1 \tag{3.1}$$

that is facet-defining for $P_k(G)$ if G_π is 2-connected. OP-inequalities (3.1) include the cut-inequalities for which G_π is a set of parallel edges. Define the polyhedron $\bar{P}_k(G)$, where

$$\bar{P}_k(G) = \{x \in R_+^E \mid x \text{ satisfies the OP-inequalities (3.1) for all partitions } \pi\}. \tag{3.2}$$

The main result of this section is the following.

Theorem 3.1. *If G is an outerplanar graph, then $\bar{P}_k(G) = P_k(G)$.*

Since each inequality defining $\bar{P}_k(G)$ is valid for $P_k(G)$, it follows that $P_k(G) \subseteq \bar{P}_k(G)$. Furthermore, each integer vertex of $\bar{P}_k(G)$ is also a vertex of $P_k(G)$. Thus, if $P_k(G) \neq \bar{P}_k(G)$, then $\bar{P}_k(G)$ must contain a fractional vertex.

Let G be an outerplanar graph, minimal with respect to the property that $\bar{P}_k(G)$ has a fractional vertex; i.e., for every minor G' of G, we have $\bar{P}_k(G') = P_k(G')$. Let x be a fractional vertex of $\bar{P}_k(G)$. The proof of Theorem 3.1 uses a sequence of propositions regarding the structure of x and G. These are stated and proved in the following.

Proposition 3.1. *If x and G are as described above, $x(e) > 0$ for all $e \in E$.*

Proof. To the contrary, assume that $x(\bar{e}) = 0$. Define $G' = (V, E - \{\bar{e}\})$. Let x' be the restriction of x to $E - \{\bar{e}\}$. x' is a fractional vertex of $\bar{P}_k(G')$, contradicting the minimality of G. □

Proposition 3.2. *The graph G is 2-connected.*

Proof. To the contrary, assume that G is not 2-connected. Let $\bar{G} = (\bar{V}, \bar{E})$ be any maximal 2-connected component of G such that there exists $\bar{e} \in \bar{E}$ with $x(\bar{e})$ fractional. Let $\bar{x}$ be the restriction of x to $\bar{E}$. $\bar{x}$ is a fractional vertex of $\bar{P}_k(\bar{G})$, since each k-edge-connected subgraph of $\bar{G}$ is the restriction of a k-edge-connected subgraph of G to $\bar{E}$. This contradicts the minimality of G. □

PROPOSITION 3.3. *Assume that degree* $(u) = 2$ *for some node* $u \in V$. *Let* $e_1 = (u, v)$ *and* $e_2 = (u, w)$ *be the two edges incident to u. Assume that* $x(e_1) \geq x(e_2)$. *Then* $x(e_1) = \lceil k/2 \rceil$ *and* $x(e_2) \in \{\lceil k/2 \rceil, \lfloor k/2 \rfloor\}$.

Proof. There are two possible cases: one where the edge $e_3 = (v, w)$ is present and the other where it is absent. Our reduction in the first case contracts e_1 and deletes e_2 and, in the second case, contracts e_1 (or e_2) to get a minor of G. To shorten the proof, we assume the edge $e_3 = (v, w)$ to be present. If this is not the case, we set $x(e_3) = 0$. The reduction we use is to contract e_1 and delete e_2, obtaining a minor of the original graph. The artifact of adding e_3 with $x(e_3) = 0$, if it is missing, allows us to resolve both cases simultaneously.

We first show that $x(e_1) \leq \lceil k/2 \rceil$. To the contrary, assume that $x(e_1) > \lceil k/2 \rceil$. We claim that there can be no OP-inequality (3.1), with $|V_\pi| \geq 3$ and $e_1 \in E_\pi$, satisfied with equality by x. To the contrary, assume that there exists a contraction minor $G_\pi = (V_\pi, E_\pi)$ of G with $|V_\pi| \geq 3$ and $e_1 \in E_\pi$ such that

$$\sum_{e \in E_\pi} x(e) = |V_\pi| \lceil k/2 \rceil - 1.$$

Let $G_{\pi 1} = (V_{\pi 1}, E_{\pi 1})$ be the graph obtained from G_π by contracting e_1 and deleting all edges parallel to e_1. Since $x(e_1) > \lceil k/2 \rceil$, we have

$$\sum_{e \in E_{\pi 1}} x(e) < |V_{\pi 1}| \lceil k/2 \rceil - 1.$$

Thus x violates the OP-inequality defined by $G_{\pi 1}$, contradicting the assumption that $x \in \bar{P}_k(G)$.

Contract the edge e_1 and delete e_2 to get the graph $G' = (V', E')$. Define x', where

$$x'(e) = \begin{cases} x(e) & \text{for } e \in E' - \{e_3\}, \\ x(e_2) + x(e_3) & \text{for } e = e_3. \end{cases}$$

If x' is a vertex of $\bar{P}_k(G')$, we contradict the minimality of G. Thus x' is not a vertex of $\bar{P}_k(G')$. This implies that there exist x'_1 and x'_2 in $\bar{P}_k(G')$ such that $x' = (x'_1 + x'_2)/2$. Given $\varepsilon > 0$, we can choose x_1, x_2 such that $|x'_i(e) - x'(e)| \leq \varepsilon$ for all $e \in E'$, $i = 1, 2$. There are two possible cases: (i) $x(e_2) < x(e_1)$ and (ii) $x(e_2) = x(e_1)$. First, consider the case where $x(e_2) < x(e_1)$. Define x_i, $i = 1, 2$, where

$$x_i(e) = \begin{cases} x(e_1) - \{x'_i(e_3) - x'(e_3)\} & \text{for } e = e_1, \\ x(e_2) + \{x'_i(e_3) - x'(e_3)\} & \text{for } e = e_2, \\ x(e_3) & \text{for } e = e_3, \\ x'_i(e) & \text{for } e \in E - \{e_1, e_2, e_3\}. \end{cases}$$

Note that $0 < x(e_2) < x(e_1)$ and that no OP-inequality (3.1) with $|V_\pi| \geq 3$ and $e_1 \in E_\pi$ is satisfied with equality by x. Thus the only OP-inequality with e_1 in its support, satisfied with equality by x, is $x(e_1) + x(e_2) \geq k$. Thus, for $\varepsilon > 0$ and small enough, $x_1, x_2 \in \bar{P}_k(G)$. Since $x = (x_1 + x_2)/2$, this contradicts the assumption that x is a vertex of $\bar{P}_k(G)$. This rules out $x(e_2) < x(e_1)$.

If $x(e_2) = x(e_1)$, define x_1 and x_2, where

$$x_i(e) = \begin{cases} x(e) + \{x_i'(e_3) - x'(e_3)\} & \text{for } e \in \{e_1, e_2\}, \\ x(e_3) & \text{for } e = e_3, \\ x_i'(e) & \text{for } e \in E - \{e_1, e_2, e_3\}. \end{cases}$$

Since $x(e_1) = x(e_2) > \lceil k/2 \rceil$, no OP-inequality satisfied with equality by x has support in both e_1 and e_2. Thus, for $\varepsilon > 0$ and small enough, $x_1, x_2 \in \bar{P}_k(G)$. Since $x = (x_1 + x_2)/2$, this contradicts the assumption that x is a vertex of $\bar{P}_k(G)$. This also rules out $x(e_2) = x(e_1)$. Thus we have proved that $x(e_1) \leq \lceil k/2 \rceil$.

Now we show that $x(e_1) = \lceil k/2 \rceil$. To the contrary, assume that $x(e_1) < \lceil k/2 \rceil$. Thus $x(e_2) < \lceil k/2 \rceil$. We now show that there exists no OP-inequality defined by $G_\pi = (V_\pi, E_\pi)$, where $e_2 \in E_\pi$, $e_1 \notin E_\pi$, and

$$\sum_{e \in E_\pi} x_e = |V_\pi| \lceil k/2 \rceil - 1.$$

To the contrary, assume that such an inequality exists. Consider the graph $G_{\pi 1} = (V_{\pi 1}, E_{\pi 1})$, obtained by expanding the edge e_1. $G_{\pi 1}$ is also a contraction minor of G and defines an OP-inequality. $V_{\pi 1} = V_\pi \cup \{u\}$ and $E_{\pi 1} = E_\pi \cup \{e_1\}$. Since $x(e_1) < \lceil k/2 \rceil$ and $|V_{\pi 1}| = V_\pi + 1$, we have

$$\sum_{e \in E_\pi} x_e < |V_{\pi 1}| \lceil k/2 \rceil - 1.$$

This contradicts the assumption that $x \in \bar{P}_k(G)$. Thus there can be no OP-inequality satisfied with equality by x where $e_2 \in E_\pi$, $e_1 \notin E_\pi$. A similar argument shows that there can be no OP-inequality satisfied with equality by x where $e_1 \in E_\pi$, $e_2 \notin E_\pi$. The two statements together contradict the assumption that x is a vertex of $\bar{P}_k(G)$. Thus $x(e_1) = \lceil k/2 \rceil$.

This implies that $\lfloor k/2 \rfloor \leq x(e_2) \leq \lceil k/2 \rceil$. We now show that $x(e_2) \in \{\lceil k/2 \rceil, \lfloor k/2 \rfloor\}$. To the contrary, assume that $\lfloor k/2 \rfloor < x(e_2) < \lceil k/2 \rceil$, yielding $x(e_1) + x(e_2) > k$. Contract e_1 and delete e_2 to get $G' = (V', E')$. Define x, where

$$x' = \begin{cases} x(e_2) + x(e_3) & \text{for } e = e_3, \\ x(e) & \text{for } e \in E' - \{e_3\}. \end{cases}$$

If x' is a vertex of $\bar{P}_k(G')$, we contradict the minimality of G. Thus x' is not a vertex of $\bar{P}_k(G')$. This implies that there exist x_1', x_2' in $\bar{P}_k(G)$ such that $x' = (x_1' + x_2')/2$. Given $\varepsilon > 0$, we can choose x_1' and x_2' such that $|x_i'(e) - x'(e)| \leq \varepsilon$ for all $e \in E$. Define x_i for $i = 1, 2$, where

$$x_i(e) = \begin{cases} x(e_1) & \text{for } e = e_1, \\ x(e_2) + \{x_i'(e_3) - x'(e_3)\} & \text{for } e = e_2, \\ x(e_3) & \text{for } e = e_3, \\ x_i'(e) & \text{for } e \in E - \{e_1, e_2, e_3\}. \end{cases}$$

For ε small enough, $x_i(e_1) + x_i(e_2) \geq k$, $i = 1, 2$. Thus $x_1, x_2 \in \bar{P}_k(G)$. Since $x = (x_1 + x_2)/2$, this contradicts the assumption that x is a vertex of $\bar{P}_k(G)$. Thus $x(e_2) \in \{\lceil k/2 \rceil, \lfloor k/2 \rfloor\}$. □

Proof of Theorem 3.1. By Proposition 3.1, G cannot contain Fig. 2.1(a). Now consider the case where G contains Fig. 2.1(b). Since x is a vertex of $\bar{P}_k(G)$, either $x(e_1) = 0$ or $x(e_2) = 0$, since, in each OP-inequality, $x(e_1)$ and $x(e_2)$ have the same coefficient. This contradicts Proposition 3.2. Thus we can assume that G is a 2-connected outerplanar graph with no parallel edges.

Order the nodes of V as $\{1, \ldots, |V|\}$ along the outer face. An edge of the form (i, j), where $j \notin \{i - 1, i + 1\}$, is called a *chord*. First, consider the case where G does not contain any chord. Thus each node $i \in V$ has degree 2. From Proposition 3.3, $x(e) = \lceil k/2 \rceil$ for $e \in E - \{\bar{e}\}$ and $x(\bar{e}) = \lfloor k/2 \rfloor$ for some edge $\bar{e} \in E$. Thus $\bar{P}_k(G)$ has only integer vertices, contradicting our assumption that x is a fractional vertex of $\bar{P}_k(G)$.

Thus G must contain a chord $\bar{e} = (i, j)$. Since G is outerplanar, we can choose a chord (i, j), where $j \geq i$ and degree$(s) = 2$ for $s \in \{i + 1, \ldots, j - 1\}$. Define $\{e_s = (s, s + 1), s = 1, \ldots, |V|)$ and the edge set $\bar{E} = \{e_s, s = i, i + 1, \ldots, j - 1\}$. From our previous results, $x(e) = \lceil k/2 \rceil$ for $e \in \bar{E} - \{\tilde{e}\}$ and $x(\tilde{e}) \in \{\lfloor k/2 \rfloor, \lceil k/2 \rceil\}$ for some edge $\tilde{e} \in \bar{E}$. Form the graph $G' = (V', E')$ by contracting the edges $\bar{E} - \{\tilde{e}\}$ and deleting $\tilde{e}$. Define x', where

$$x' = \begin{cases} x(\bar{e}) + x(\tilde{e}) & \text{for } e = \bar{e}, \\ x(e) & \text{for } e \in E' - \{\bar{e}\}. \end{cases}$$

Clearly, $x' \in \bar{P}_k(G')$. If x' is a vertex of $\bar{P}_k(G')$, we contradict the minimality of G. Thus x' is not a vertex of $\bar{P}_k(G')$. This implies that there exist x'_1 and x'_2 in $\bar{P}_k(G')$ such that $x' = (x'_1 + x'_2)/2$. Given $\varepsilon > 0$, we can choose x'_1 and x'_2 such that $|x'_1(e) - x'(e)| \leq \varepsilon$ for all $e \in E$, $i = 1, 2$.

Define x_i, $i = 1, 2$, where

$$x_i(e) = \begin{cases} x(e) & \text{for } e \in \bar{E}, \\ x(\bar{e}) + x'_i(\bar{e}) - x'(\bar{e}) & \text{for } e = \bar{e}, \\ x'_i(e) & \text{for } e \in E' - \{\bar{e}\}. \end{cases}$$

By Proposition 3.1, $x(\bar{e}) > 0$. Thus, for ε small enough, $x_1, x_2 \in \bar{P}_k(G)$ and $x = (x_1 + x_2)/2$. This contradicts the assumption that x is a vertex of $\bar{P}_k(G)$. This shows that $\bar{P}_k(G)$ does not contain any fractional vertex. The result thus follows. □

If $\bar{P}_k(G)$ is described using only OP-inequalities as in (3.2), then $\bar{P}_k(G) \neq P_k(G)$, even if G is the complete graph on four nodes (K_4), as shown in Fig. 3.1. $x(e_i) = 1.5$, $i = 1, 2, 5, 6$; $x(e_4) = 0$; $x(e_3) = 0.5$ is a fractional vertex of $\bar{P}_3(G)$. This shows that Theorem 3.1 cannot be extended to graphs with a K_4 minor. This fractional vertex is cut off by the LOP-inequality

$$x(e_1) + x(e_2) + x(e_3) + x(e_5) + x(e_6) + 2x(e_4) \geq 7.$$

However, we do not know of any series-parallel graphs (graphs with no K_4 minor) for which $\bar{P}_k(G) \neq P_k(G)$. Thus we make the following conjecture.

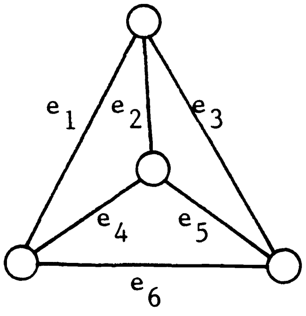


FIG. 3.1

CONJECTURE 3.1. *$\bar{P}_k(G) = P_k(G)$ if k is odd and G is a series-parallel graph.*

It would also be interesting to characterize those graphs for which $P_k(G)$ is completely defined by LOP-inequalities. Results in §4 give examples of graphs for which this is not the case.

4. Odd-wheel inequalities. In this section, we describe a family of facet-defining inequalities for $P_k(G)$. Given an odd integer $s \geq 3$ and integers $n_i \geq 2$, $i \in \{1, \ldots, s\}$, define the *odd-wheel configuration* $W_s = (V_s, E_s)$, where

$$V_s = \{u\} \cup \{v_j^i \mid j \in \{1, \ldots, n_i\},\ i \in \{1, \ldots, s\}\}.$$

Define $v_0^i = u$ for $i \in \{1, \ldots, s\}$. The edge set E_s is given by

$$E_s = \{(v_j^i, v_{j+1}^i) \mid j \in \{0, \ldots, n_i - 1\},\ i \in \{i, \ldots, s\}\} \cup \{(v_{n_i}^i, v_{n_{i+1}}^{i+1}) \mid i \in \{1, \ldots, s\}\}.$$

For $s = 3$ and $n_i = 2$, $i \in \{1, 2, 3\}$, W_3 is shown in Fig. 4.1. The precise inequality we define on W_s depends upon the value of k. There are three distinct sets of values of k we consider: $k \in \{4r - 2, 4r - 1, 4r + 1, r \in \{1, 2, \ldots\}\}$. We give proofs only for the case where $k = 4r + 1$. For the other two cases, we simply state the inequality. Detailed proofs are contained in Chopra [3].

Consider the odd-wheel configuration $W_s = (V_s, E_s)$, where $k = 4r + 1$ and $n_i \geq 2$, $i \in \{1, \ldots, s\}$. Define the edges $e_{ij} = (v_j^i, v_{j+1}^i)$ for $i \in \{1, \ldots, s\}$ and $j \in \{0, \ldots, n - 1\}$ and $e_i = (v_{n_i}^i, v_{n_{i+1}}^{i+1})$ for $i \in \{1, \ldots, s\}$. The odd-wheel inequality is given by

$$\sum_{e \in E_s} a(e)x(e) \geq a_0, \tag{4.1}$$

where $a(e) = 1$ for all $e \in E_s$ and $a_0 = s(r - 1) + (2r + 1) \sum_{i=1}^{s} n_i + \lceil s/2 \rceil$.

LEMMA 4.1. *The odd-wheel inequality* (4.1) *is valid for $P_k(W_s)$, $k = 4r + 1$.*

Proof. For $m \in \{1, \ldots, s\}$, consider the OP-inequalities

$$\sum_{j \in \{0, \ldots, n_m - 1\}} x(e_{mj}) + x(e_{m-1}) + x(e_m) \geq (n_m + 1)(2r + 1) - 1 \tag{4.2}$$

and

$$\sum_{j \in \{0, \ldots, n_m - 1\}} x(e_{mj}) \geq n_m(2r + 1) - 1. \tag{4.3}$$

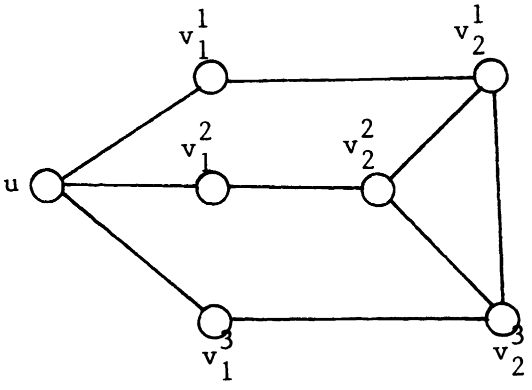


FIG. 4.1

Adding (4.2) and (4.3) for all $m \in \{1, \dots, s\}$ and dividing by 2 gives

$$(4.4) \qquad \sum_{e \in E_s} x(e) \geq s(r-1) + (2r+1) \sum_{i=1}^{s} n_i + (s/2).$$

On rounding (4.4), we obtain (4.1). □

THEOREM 4.1. *The odd-wheel inequality is facet-defining for $P_k(W_s)$, $k = 4r + 1$.*

Proof. Let $\alpha x \geq \alpha_0$ be any valid inequality for $P_k(W_s)$ such that

$$\{x \in P_k(W_s) \mid x \text{ satisfies (4.1) with equality}\} \subseteq \{x \mid \alpha x = \alpha_0\}.$$

Given $m \in \{1, \dots, s\}$ and $l_i \in \{0, \dots, n_i - 1\}$ for $i = 1, \dots, s$, define x_m^1, where

$$x_m^1(e) = \begin{cases} 2r & \text{for } e = e_{ij}, \quad i \in \{1, \dots, s\}, \quad j = l_i, \\ 2r+1 & \text{for } e = e_{ij}, \quad i \in \{1, \dots, s\}, \quad j \neq l_i, \\ r+1 & \text{for } e = e_i, \quad i \in \{m, m+2, \dots, m+s-1\}, \\ r & \text{for } e = e_i, \quad i \in \{m+1, m+3, \dots, m+s-2\}. \end{cases}$$

Define x_m^2, where

$$x_m^2(e) = \begin{cases} 2r+1 & \text{for } e = e_{m+1,j}, \quad j = l_{m+1}, \\ r & \text{for } e = e_m, \\ x_m^1(e) & \text{otherwise}. \end{cases}$$

x_m^1 and x_m^2 satisfy (4.2) at equality for $m \in \{1, \dots, s\}$ and $l_i \in \{0, \dots, n_i - 1\}$. Comparing x_m^1 for $l_i = j$ and $l_i = k$ gives

$$(4.5) \qquad \alpha(e_{ij}) = \alpha(e_{ik}) = b_i \quad \text{for } i \in \{1, \dots, s\}, \quad j \neq k \in \{0, \dots, n_i - 1\}.$$

Comparing x_m^1 and x_m^2, we have

$$(4.6) \qquad \alpha(e_m) = b_{m+1} \quad \text{for } m \in \{1, \dots, s\}.$$

Comparing x_m^1 and x_{m+2}^1, we have

$$(4.7) \qquad \alpha(e_m) = \alpha(e_{m+1}) \quad \text{for } m \in \{1, \dots, s\}.$$

From (4.5)–(4.7), we have $\alpha(e) = b$ for $e \in E_s$. This shows that $\alpha x \geq \alpha_0$ is a multiple of (4.1). The result thus follows. □

In the case where $k = 4r - 1$, the odd-wheel inequality is given by (4.1), where

$$a(e) = \begin{cases} 1/(n_i - 1) & \text{for } e = e_{ij}, i \in \{1, \dots, s\}, j \in \{0, \dots, n_i - 1\}, \\ 1 + 1/(n_i - 1) + 1/(n_{i+1} - 1) & \text{for } e = e_i, i \in \{1, \dots, s\} \end{cases}$$

and $a_0 = (3r-1)s + \lceil s/2 \rceil + (4r-1) \sum_{i=1}^{s} (1/(n_i - 1))$.

In the case where $k = 4r - 2$, we restrict $n_i = 2$ for $i \in \{1, \dots, s\}$ in W_s. The odd-wheel inequality is once again given by (4.1), where

$$a(e) = \begin{cases} \frac{1}{2} & \text{for } e = (v_j^i, v_{j+1}^i), \quad j \in \{0, 1\}, \\ 1 & \text{for } e = (v_2^i, v_2^{i+1}), \quad i \in \{1, \dots, s\} \end{cases}$$

and $a_0 = s(3r-2) + \lceil s/2 \rceil$.

For the case where $k = 2$, a similar inequality was considered by Mahjoub [11] and generalized by Stoer [14] and Boyd and Hao [2].

4.1. High edge-connectivity network design. In this section, we relate facet-defining inequalities of $P_k(G)$ to facet-defining inequalities of polyhedra associated with high edge-connectivity networks. The problem of designing networks with high edge-connectivity was first introduced by Grötschel and Monma [10] and was studied in detail from a polyhedral point of view by Stoer [14]. Both restrict attention to the case where an edge can be used at most once. In the following description, we use the terminology of Grötschel and Monma [10]. Consider an undirected graph $G = (V, E)$ with nonnegative edge-weights w_e. For each node $v \in V$, a nonnegative integer r_v is specified. For each pair of nodes u, v in V, define $r_{uv} = \min \{r_u, r_v\}$. A family of edges F is said to satisfy the edge-connectivity requirements if between every pair of nodes u, v in V, $G[F]$ contains at least r_{uv} edge disjoint paths. The problem is to find a minimum-weight family F that satisfies the edge-connectivity requirements.

Define the associated polyhedron

$$\mathrm{PCON}(G, r) = \mathrm{conv}\ \{x(F) \mid F \text{ satisfies edge-connectivity requirements}\}.$$

Given $U \subseteq V$, define $r(U) = \max \{r_v \mid v \in U\}$. Consider a partition $\pi = (V_i,\ i = 1, \dots, k)$ of the node set V, where each subset V_i induces a connected subgraph. As described earlier, $G_\pi = (V_\pi, E_\pi)$, the graph obtained by shrinking each subset V_i into a single node, is a contraction minor of G. Assume that $k_1 = \max \{r(V_i),\ i = 1, \dots, l\}$ and $r(V_l) = k_1$. Further assume that $r(V_i) = k$ for $i = 1, \dots, l - 1$. We can relate facet-defining inequalities of $P_k(G_\pi)$ and PCON (G, r) by the following result. The proof follows from Theorem 2.4.

THEOREM 4.2. *If the inequality* $\sum_{e \in E_\pi} a_e x_e \ge a_0$ *is facet-defining* (*valid*) *for* $P_k(G_\pi)$, *then* $\sum_{e \in E} a_e x_e \ge a_0$ *is facet-defining* (*valid*) *for* PCON (G, r), *where* $a_e = 0$ *for* $e \in E - E_\pi$.

5. Some open problems. In this paper, we introduce the OP-inequalities and show that these completely describe $P_k(G)$ for k odd if G is outerplanar. One open problem is to see if the same result holds when G is series-parallel. Another open problem is to characterize those graphs G for which $P_k(G)$ is completely described by the cut-inequalities where $k = 2^r$, $r = 2, 3, \dots$. This would generalize the result of Cornuéjols, Fonlupt, and Naddef [5], who characterized the case where $r = 1$.

Acknowledgments. I thank Bruce Gamble for several valuable discussions in the course of this work. I also thank the referees for a very careful reading of the paper and several valuable suggestions. Mechthild Stoer suggested inequality (2.7) and Theorem 2.6.

REFERENCES

[1] D. BIENSTOCK, E. F. BRICKELL, AND C. L. MONMA, *On the structure of minimum-weight k-connected spanning networks*, SIAM J. Discrete Math., 3 (1990), pp. 320–329.

[2] S. C. BOYD AND T. HAO, *An Integer Polytope Related to the Design of Survivable Communication Networks*, Tech. Report TR-91-18, Computer Science Department, University of Ottawa, Ottawa, Canada, 1991.

[3] S. CHOPRA, *The k-edge Connected Spanning Subgraph Polyhedron*, preprint, Northwestern University, Evanston, IL, 1991.

[4] N. CHRISTOFIDES AND C. A. WHITLOCK, *Network synthesis with connectivity constraints—A survey*, in Operational Research '81, J. P. Brans, ed., North-Holland, Amsterdam, pp. 705–723.

[5] G. CORNUÉJOLS, F. FONLUPT, AND D. NADDEF, *The traveling salesman problem on a graph and some related integer polyhedra*, Math. Programming, 33 (1985), pp. 1–27.

[6] H. GAN AND E. L. JOHNSON, *Four problems on graphs with excluded minors*, Mathematical Programming, 45 (1989), pp. 311–330.

[7] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman and Co., San Francisco.

[8] M. X. GOEMANS, *Analysis of linear programming relaxations for a class of connectivity problems*, MIT Research Report OR 233-90, Cambridge, MA, 1990.

[9] M. X. GOEMANS AND D. J. BERTSIMAS, *Survivable networks, linear programming relaxations and the parsimonious property*, MIT Research Report OR 216-90, Cambridge, MA, 1990.

[10] M. GRÖTSCHEL AND C. L. MONMA, *Integer polyhedra associated with certain network design problems with connectivity constraints*, SIAM J. Discrete Math., 3 (1990), pp. 502–523.

[11] A. R. MAHJOUB, *Two edge connected spanning subgraphs and polyhedra*, Tech. Report No. 88520-OR, Institut für Ökonometrie und Operations Research, Bonn, Germany, 1988.

[12] C. L. MONMA, B. S. MUNSON, AND W. R. PULLEYBLANK, *Minimum weight two-connected spanning networks*, Math. Programming, 46 (1990), pp. 153–172.

[13] W. R. PULLEYBLANK, *Polyhedral combinatorics*, in Handbooks in OR & MS, Vol. 1, G. L. Nemhauser et al., eds., North-Holland, Amsterdam, 1989, pp. 371–446.

[14] M. STOER, *Design of Survivable Networks*, Ph.D. thesis, Universität Augsburg, Germany, 1991.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 260-274, May 1994

CONSTRUCTING SMALL SAMPLE SPACES SATISFYING GIVEN CONSTRAINTS*

DAPHNE KOLLER† AND NIMROD MEGIDDO‡

Abstract. The subject of this paper is finding small sample spaces for joint distributions of n discrete random variables. Such distributions are often only required to obey a certain limited set of constraints of the form Pr (Event) = π. It is shown that the problem of deciding whether there exists any distribution satisfying a given set of constraints is NP-hard. However, if the constraints are consistent, then there exists a distribution satisfying them, which is supported by a "small" sample space (one whose cardinality is equal to the number of constraints). For the important case of *independence constraints*, where the constraints have a certain form and are consistent with a joint distribution of *independent* random variables, a small sample space can be constructed in polynomial time. This last result can be used to derandomize algorithms; this is demonstrated by an application to the problem of finding large independent sets in sparse hypergraphs.

Key words. discrete probability distribution, linear programming, algorithm, sample space, derandomization, hypergraph, independent set, probabilistic constraint satisfaction, independent random variables

AMS subject classifications. 60E05, 68R99

1. Introduction. The probabilistic method of proving existence of combinatorial objects has been very successful (see, for example, Raghavan [15] and Spencer [17]). The underlying idea is as follows. Consider a finite set Ω whose elements are classified as "good" and "bad." Suppose that we wish to prove existence of at least one "good" element within Ω. The proof proceeds by constructing a probability distribution f over Ω and showing that the probability of picking a good element is positive. Probabilistic proofs often yield randomized algorithms for constructing a good element. In particular, many randomized algorithms are a special case of this technique, where the "good" elements are those sequences of random bits leading to a good answer.

It is often desirable to replace the probabilistic construction by a deterministic one, or to *derandomize* an algorithm. Obviously, this can be done by completely enumerating the sample space Ω until a good element is found.[1] Unfortunately, the sample space is typically exponential in the size of the problem; for example, the sample space of n independent random bits[2] contains 2^n points.

Let $X_1, \ldots, X_n$ be discrete random variables with a finite range. For simplicity, we assume that $X_1, \ldots, X_n$ all have the same range $\{0, \ldots, r-1\}$ (although not necessarily the same distribution). Our constructions can easily be extended to variables with different

* Received by the editors March 23, 1992; accepted for publication (in revised form) April 30, 1993. A preliminary version of this paper appeared in Proceedings of the 25th Annual ACM Symposium on Theory of Computing, 1993. This research was supported in part by Office of Naval Research contract N00014-91-C-0026 and by Air Force Office of Scientific Research contract F49620-91-C-0080. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes.

† Computer Science Division, University of California, Berkeley, California 94720, and IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120 (daphne@cs.berkeley.edu).

‡ IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120, and School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel (megiddo@almaden.ibm.com).

[1] This can be done if we assume that good elements are easy to recognize. In decision problems, this is usually the case. In optimization problems, we may be able to prove that a random element is optimal or close to optimal with a certain probability. In those cases, although we may not be able to tell by looking at an element if it is "good," we can often compare elements and decide which is "better." We can therefore derandomize such an algorithm by enumerating the sample space and choosing the "best" element in it. The techniques of this paper also apply to problems of this type.

[2] We use the term *random bits* to denote binary-valued, uniformly distributed random variables.

ranges. The *probability space* associated with these variables is $\Omega = \{0, \ldots, r-1\}^n$. A *distribution* is a map $f\colon \Omega \to [0, 1]$ such that $\sum_{x \in \Omega} f(x) = 1$. We define the set $S(f) = \{x \in \Omega \mid f(x) > 0\}$ to be the *sample space* of f.

Given a distribution f involved in a probabilistic proof, only the points in $S(f)$ must be considered in our search for a good point in Ω. Moreover, it suffices to search any subset of $S(f)$ that is guaranteed to contain a good point for each possible input. Adleman [1] shows that, for any distribution f used in an algorithm in RP, there exists a space $S' \subseteq S(f)$ of polynomial size that contains a good point for every possible input. The proof of this fact is not constructive and therefore cannot be used to derandomize algorithms.

A common technique for constructing a feasible search-space is to find a different distribution with a "small" (polynomial) sample space that can be searched exhaustively, as outlined above. The new distribution must agree with the original one sufficiently so that the correctness proof of the algorithm remains valid. The correctness proof often relies on certain assumptions about the distribution; that is, the distribution is assumed to satisfy certain constraints. A *constraint* is an equality of the form

$$\Pr(Q) = \sum_{x \in Q} f(x) = \pi,$$

where $Q \subseteq \Omega$ is an *event* and $0 \leq \pi \leq 1$. If the randomness requirements of an algorithm are completely describable as a set of constraints and if the new distribution satisfies all of them, then the algorithm remains correct under the new distribution; no new analysis is needed. In other cases, the new distribution may only approximately satisfy the constraints, and it is necessary to verify that the analysis still holds.

The original distribution is almost always constructed based on *independent* random variables $X_1, \ldots, X_n$. Thus, all the constraints are satisfied by such a distribution. In many cases, however, full independence is not necessary. In particular, quite often the constraints are satisfied by a *d-wise independent distribution*—a distribution where each *neighborhood* of d variables behaves as if it were independent. That is, it suffices for the distribution to satisfy the *independence constraints* which state that every event defined over a neighborhood of size d has the same probability as if the variables were independent. Most of the previous work has focused on constructing approximations to such distributions.

Joffe [10] first demonstrated a construction of a joint distribution of n d-wise independent, uniformly distributed random variables with a sample space of cardinality $O((2n)^d)$. Luby [12] and Alon, Babai, and Itai [3] show how Joffe's construction can be generalized to allow for nonuniform distributions using sample spaces of essentially the same cardinality. In many cases, the resulting distributions only approximately satisfy the required constraints; that is, the distributions are d-wise independent, but the probabilities $\Pr(X_i = b)$ may differ from the corresponding probabilities in the original distribution.[3] These constructions result in sample spaces of polynomial size for any fixed d. Chor et al. [8] showed that any sample space of n d-wise independent random bits has cardinality $\Omega(n^{\lfloor d/2 \rfloor})$. Thus, these constructions are close to optimal in this case. Moreover, sample spaces of polynomial size exist for d-wise independent distributions only if d is fixed.

Naor and Naor [14] showed how to circumvent this lower bound by observing that *ε-independent* (or *nearly independent*) distributions often suffice. In other words, it suffices

[3] In fact, these distributions all have a sample space of cardinality $O(p^d)$ for some prime $p \geq n$. The approximation is better for larger p's.

that the independence constraints for the neighborhoods of size d be satisfied to within ε. We point out that this is also a form of approximation, as defined above. Naor and Naor demonstrated a construction of sample spaces for ε-independent distributions over random bits, whose size is polynomial in n and in $1/\varepsilon$. These constructions are polynomial for $\varepsilon = 1/\text{poly}(n)$; for such values of ε, the ε-independence constraints are meaningful[4] for neighborhoods of size up to $O(\log n)$. Therefore, we obtain a polynomial-size sample space that is nearly d-wise independent for $d = O(\log n)$ (as compared to the lower bound of $\Omega(n^{\log n})$ for truly d-wise independent sample spaces). Simplified constructions with similar properties were provided by Alon et al. [4]. Azar, Motwani, and Naor [5] later generalized these techniques to uniform distributions over nonbinary random variables. Finally, Even et al. [9] presented constructions for nearly independent distributions over nonuniform nonbinary random variables.

A different type of technique was introduced by Berger and Rompel [7] and by Motwani, Naor, and Naor [13]. This technique can be used to derandomize certain RNC algorithms, where d, the degree of independence required, is polylogarithmic in n. The technique works, however, only for certain types of problems and does not seem to generalize to larger degrees of independence.

Schulman [16] took a different approach toward the construction of sample spaces that require $O(\log n)$-wise independence. He observed that, in many cases, only certain d-neighborhoods (sets of d variables) must be independent. Schulman constructs sample spaces satisfying this property, whose size is 2^d times the grestest number of neighborhoods to which any variable belongs. In particular, for polynomially many neighborhoods, each of size $O(\log n)$, this construction results in a polynomial-size sample space. His construction works only for random bits and is polynomial for a maximum neighborhood size $O(\log n)$.

To improve on these results, we view the problem from a somewhat different perspective. Rather than placing upper bounds on the degree of independence required by the algorithm, we examine the set of precise constraints that are required for the algorithm to work. We then construct a distribution satisfying these constraints exactly. In many cases, this approach will yield a much smaller sample space.

We begin by showing a connection between the number of constraints and the size of the resulting sample space. We show in §2 that, for any set $\mathscr{C}$ of such constraints, if $\mathscr{C}$ is *consistent*, i.e., $\mathscr{C}$ is satisfied by some distribution f, then there exists a distribution f' also satisfying $\mathscr{C}$ such that $|S(f')| \leq |\mathscr{C}|$; that is, there exists a distribution for which the cardinality of the sample space is not more than the number of constraints. Note that f' precisely satisfies the constraints in $\mathscr{C}$, so that, if $\mathscr{C}$ represents all the assumptions about f made by a proof, the proof will also hold for f'. The proof of the existence theorem includes an algorithm for constructing f'; however, the algorithm takes exponential time and is thus not useful. We justify this exponential behavior by showing that, even for a set of very simple constraints, the problem of recognizing whether there exists a distribution satisfying them is NP-complete.

We can, however, define a type of constraint for which a small sample space can be constructed directly from the constraints in polynomial time. As we observed, the distributions that are most often used in probabilistic proofs are those where $X_1, \ldots, X_n$ are independent random variables. Such a distribution is determined entirely by the

[4] Consider a distribution over random bits and some neighborhood of size k. The "correct" probability of any event prescribing values to all the variables in this neighborhood is $1/2^k$. For $k = \log(1/\varepsilon) = \Theta(\log n)$, this probability is $\leq \varepsilon$. For larger k, all such constraints are therefore subsumed by constraints corresponding to smaller neighborhoods.

probabilities $\{p_{ib} = \Pr(X_i = b)\colon i = 1, \ldots, n;\, b = 0, \ldots, r-1\}$. In the course of such a probabilistic proof, the distribution is assumed to satisfy various constraints. Above, we observed that, in many cases, and in particular in all cases for which existing constructions work, these constraints are independence constraints. More formally, an independence constraint is one that forces the probability of a certain assignment of values to some subset of the variables to be as if the variables are independent. That is, for a fixed set of p_{ib}'s, a sequence of indices $i_1, \ldots, i_k$ in $\{1, \ldots, n\}$, and $b_1, \ldots, b_k \in \{0, \ldots, r-1\}$, the constraint

$$\left[\Pr(\{X_{i_1} = b_1, \ldots, X_{i_k} = b_k\}) = \prod_{j=1}^{k} p_{i_j b_j}\right]$$

is the independence constraint $I(Q)$ corresponding to the event[5] $Q = \{X_{i_1} = b_1, \ldots, X_{i_k} = b_k\}$. Obviously, if $X_1, \ldots, X_n$ are independent random variables, then their joint distribution satisfies all the independence constraints. Note that d-wise independence can easily be represented in terms of constraints of this type: The variables $X_1, \ldots, X_n$ are d-wise independent if and only if all the independence constraints $I(\{X_{i_1} = b_1, \ldots, X_{i_d} = b_d\})$ are satisfied, where $i_1, \ldots, i_d \in \{1, \ldots, n\}$ and $b_1, \ldots, b_d \in \{0, \ldots, r-1\}$.

Let $\mathscr{C}$ be a set of independence constraints defined using a fixed set of p_{ib}'s as above. In §3 we present the main result of this paper, which shows how to construct in strongly polynomial time a distribution satisfying $\mathscr{C}$ with a sample space of cardinality $|\mathscr{C}|$. We note that the distribution f produced by our technique is typically not the uniform distribution over $S(f)$. Therefore, we cannot in general use our construction to reduce the number of random bits required to generate the desired distribution.

Our construction has a number of advantages. First, the distributions generated always satisfy the constraints precisely. Thus, the correctness proof of the algorithm need not be modified. Moreover, the size of the sample space in all the nearly independent constructions [4], [5], [9], [14] depends polynomially on $1/\varepsilon$ (where ε is the approximation factor). Our precise construction does not have this term. Previously, precise distributions were unavailable for many interesting distributions. In particular, our approach can construct sample spaces of cardinality $O((rn)^d)$ for any set of n r-valued, d-wise independent random variables (not necessarily uniformly distributed). For fixed d, this construction requires polynomial time. It has been argued by Even et al. [9] that probability distributions over nonuniform nonbinary random variables are important. To our knowledge, this is the first technique that allows the construction of exact distributions of d-wise independent variables with arbitrary p_{ib}'s.

The main advantage of our construction is that the size of the sample space depends only on the number of constraints actually used. Except for Schulman's approach [16], all other sample spaces are limited by requiring that all neighborhoods of a particular size be independent (or nearly independent). As Schulman points out, in many cases only certain neighborhoods are ever relevant, thus enabling a further reduction in the size of the sample space. However, Schulman's approach still requires the sample space to satisfy all the independence constraints associated with the relevant neighborhoods.[6] This restricts his construction to neighborhoods of maximal size $O(\log n)$. With our construction, we can deal with neighborhoods of any size, as long as the number of relevant constraints is limited.

[5] Throughout this paper, we assume without loss of generality that $i_1 < i_2 < \cdots < i_k$ and regard this notation as a shorthand for the event $\{(x_1, \ldots, x_n)\colon x_{i_1} = b_1, \ldots, x_{i_k} = b_k\}$.

[6] Moreover, as we have observed, Schulman's construction works only for random bits.

For example, an algorithm may randomly choose edges in a graph by associating a binary random variable with each edge. An event whose probability may be relevant to the analysis of this algorithm is "no edge adjacent to a node v is chosen." Using the other approaches (even Schulman's), the neighborhood size would be the maximum degree Δ of a node in the graph; the relevant sample space would then grow as 2^{Δ}. Using our approach, there is only one event per node, resulting in a sample space of size n (the number of nodes in the graph).

In this example, the constraints depend on the edge structure of the input graph. In general, our construction depends on the specific constraints derived from the particular input. Therefore, unlike most sample space constructions, our construction cannot be prepared in advance. This property, combined with the fact that our algorithm is sequential, means that it cannot be used to derandomize parallel (RNC) algorithms.

In §4 we show an example of how our technique can be applied to derandomization of algorithms. We discuss the problem of finding a large independent set in a d-uniform hypergraph. The underlying randomized algorithm, described by Alon, Babai, and Itai [3], was derandomized in the same paper for fixed values of d. It was later derandomized also for $d = O(\text{polylog } n)$ by Berger and Rompel [7] and by Motwani, Naor, and Naor [13]. We show how this algorithm can be derandomized for any d. A sequential deterministic polynomial time solution for the large independent set problem in hypergraphs exists [2]. However, the derandomization of this algorithm using our technique serves to demonstrate its unique power.

2. Existence of small sample spaces. Let $\mathscr{C} = \{[\Pr(Q_k) = \pi_k] : k = 1, \dots, c\}$ be a set of constraints such that $[\Pr(\Omega) = 1] \in \mathscr{C}$. Henceforth, the term "polynomial" means polynomial in terms of n, r, $|\mathscr{C}|$, and the bit lengths of the π_k's.

DEFINITION 2.1. *A set $\mathscr{C}$ of constraints is* consistent *if there exists some distribution f satisfying all the members of $\mathscr{C}$.*

DEFINITION 2.2. *A distribution f that satisfies $\mathscr{C}$ is said to be* manageable *if $|S(f)| \le c = |\mathscr{C}|$.*

THEOREM 2.3. *If $\mathscr{C}$ is consistent, then $\mathscr{C}$ is satisfied by a manageable distribution.*

Proof. Let $\mathscr{C}$ be as above and recall that $c = |\mathscr{C}|$. We describe a distribution f satisfying $\mathscr{C}$ as a nonnegative solution to a set of linear equations. Let $\boldsymbol{\pi} \in \mathbb{R}^c$ denote the vector $(\pi_k)_{k=1,\dots,c}$. Recall that $\Omega = \{0, \dots, r-1\}^n$, let $m = |\Omega| = r^n$, and let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ denote the points of Ω. The variable v_l represents the probability $f(\boldsymbol{x}_l)$. Let $\boldsymbol{v}$ be the vector $(v_l)_{l=1,\dots,m}$. A constraint $\Pr(Q_k) = \pi_k$ can be represented as the linear equation

$$\sum_{l=1}^{m} a_{kl} v_l = \pi_k,$$

where

$$a_{kl} = \begin{cases} 1 & \text{if } \boldsymbol{x}_l \in Q_k, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the constraints in $\mathscr{C}$ can be represented by a system $A\boldsymbol{v} = \boldsymbol{\pi}$ of linear equations (where A is the matrix (a_{kl})). Since $\mathscr{C}$ is assumed to be consistent, there is a distribution f satisfying $\mathscr{C}$. Therefore, for $v_l = f(\boldsymbol{x}_l)$, the vector $\boldsymbol{v}$ is a nonnegative solution to this system. A classical theorem in linear programming asserts that, under these conditions, there exists a basic solution to this system. That is, there exists a vector $\boldsymbol{v}' \ge \boldsymbol{0}$ such that $A\boldsymbol{v}' = \boldsymbol{\pi}$ and the columns A_{*j} such that $v'_j > 0$ are linearly independent. Let f' be the distribution corresponding to this solution vector $\boldsymbol{v}'$. Since the number of rows in the matrix is c, the number of linearly independent columns is also at most c. Therefore, the number of positive indices in $\boldsymbol{v}'$, which is precisely $|S(f')|$, is at most $c = |\mathscr{C}|$. □

```
Algorithm 1. Reduction to basic solutions.
While $\{A_{*j} : j \in S(\boldsymbol{v})\}$ are linearly dependent:
  1. Find a nonzero vector $\boldsymbol{u} \in \mathbb{R}^m$ such that:
        $u_j = 0$ for every $j \notin S(\mathbf{v})$, and
        $A\boldsymbol{u} = \mathbf{0}$.
  2. Find some $t \in \mathbb{R}$ such that:
        $\boldsymbol{v} + t\boldsymbol{u} \geq \mathbf{0}$, and
        $v_j + tu_j = 0$ for some $j \in S(\boldsymbol{v})$.
  3. Replace $\boldsymbol{v} \leftarrow \boldsymbol{v} + t\boldsymbol{u}$.
```

This theorem can be proved constructively, based on the standard algorithm outlined above. This algorithm begins with a distribution vector $\boldsymbol{v}$ and removes points from the sample space one at a time. The removal is done while keeping all variables nonnegative, so that the truth of the equations is maintained. This results in a manageable distribution vector $\boldsymbol{v}'$. Throughout the algorithm, $S(\boldsymbol{v})$ denotes the set of indices $\{j\colon v_j > 0\}$. Intuitively, these indices represent points in the sample space of the distribution represented by $\boldsymbol{v}$.

Algorithm 1 is described in full detail by Beling and Megiddo [6]. They show that it requires $O(|S(f)| \cdot c^2)$ arithmetic operations, assuming that f is represented sparsely (so that points not in $S(f)$ need not be considered at all).[7] However, Beling and Megiddo also present a faster algorithm for the same problem, based on fast matrix multiplication. Given a matrix multiplication algorithm that multiplies two $k \times k$ matrices using $O(k^{2+\delta})$ arithmetic operations, the algorithm of Beling and Megiddo finds a basic solution in $O(c^{(3-\delta)/(2-\delta)}|S(f)|)$ arithmetic operations. Using the best algorithm known for matrix multiplication, their algorithm allows us to prove the following result.

THEOREM 2.4. *Given a distribution f in sparse representation that satisfies the constraints in $\mathscr{C}$, it is possible to construct a manageable distribution f' satisfying the same constraints using $O(|S(f)| \cdot c^{1.62})$ arithmetic operations.*

Unfortunately, the complexity of this approach is linear in $|S(f)|$, which can be as large as $m = r^n$. The algorithm is therefore exponential in n in the worst case.[8]

The exponential behavior of these algorithms can be justified by considering the problem of deciding whether a given set of constraints $\mathscr{C}$ is consistent; that is, does there exist a distribution f satisfying the constraints in $\mathscr{C}$? For arbitrary constraints, the representation of the events can be very long, causing the input size to be unreasonably large. We therefore restrict attention to *simple constraints*.

DEFINITION 2.5. *We say that a constraint* $\Pr(Q) = \pi$ *is* k-simple *if there exist* $i_1, \ldots, i_k \in \{1, \ldots, n\}$ *and* $b_1, \ldots, b_k \in \{0, \ldots, r-1\}$ *such that* $Q = \{X_{i_1} = b_1, \ldots, X_{i_k} = b_k\}$. *A constraint is* simple *if it is k-simple for some k.*

Note that the natural representation of the event as a simple constraint requires space that is at most linear in n, whereas the number of points in the event is often exponential in n (for example, a 1-simple constraint contains r^{n-1} points). We assume throughout that simple constraints are represented compactly (in linear space). Under this assumption, we can show that the consistency problem is NP-hard, even when restricted to 2-simple constraints over binary-valued random variables.

[7] If f is not represented sparsely, it obviously requires exponential time simply to read it in.

[8] The manageable distribution can also be computed directly from the constraints using a linear programming algorithm that computes basic solutions. The running time of such an algorithm will also be exponential in n.

PROPOSITION 2.6. *The problem of recognizing whether a set $\mathscr{C}$ of 2-simple constraints is consistent is* NP-*hard, even if the variables constrained by $\mathscr{C}$ are binary-valued.*

Proof. The proof uses a reduction from the 3-colorability problem: Given a graph $G = (V, E)$, decide if there exists a legal coloring γ: $V \to \{1, 2, 3\}$. Let G be a graph, and assume that $V = \{v_1, \ldots, v_n\}$. We define a set of $3n$ binary-valued variables $\{X_{i,1}, X_{i,2}, X_{i,3} : i = 1, \ldots, n\}$. Intuitively, we would like it to be the case that $\gamma(v_i) = c$ if and only if $X_{i,c} = 1$ and $X_{i,b} = 0$ for $b \neq c$; for example, $\gamma(v_i) = 2$ if and only if $X_{i,1} = X_{i,3} = 0$ and $X_{i,2} = 1$. We construct $\mathscr{C}$ so that the constraints enforce this relationship. The set $\mathscr{C}$ contains constraints of two types.

- For each $i = 1, \ldots, n$ and $b \neq c \in \{1, 2, 3\}$, $\mathscr{C}$ contains the constraints

$$\Pr(\{X_{i,b} = 0, X_{i,c} = 0\}) = \tfrac{1}{3},$$

$$\Pr(\{X_{i,b} = 0, X_{i,c} = 1\}) = \tfrac{1}{3},$$

$$\Pr(\{X_{i,b} = 1, X_{i,c} = 0\}) = \tfrac{1}{3}.$$

Intuitively, these disallow illegal colorings, where the same node gets two colors.

- For each $(v_i, v_j) \in E$ and each $b \in \{1, 2, 3\}$, $\mathscr{C}$ contains the constraints

$$\Pr(\{X_{i,b} = 0, X_{j,b} = 0\}) = \tfrac{1}{3},$$

$$\Pr(\{X_{i,b} = 0, X_{j,b} = 1\}) = \tfrac{1}{3},$$

$$\Pr(\{X_{i,b} = 1, X_{j,b} = 0\}) = \tfrac{1}{3}.$$

Intuitively, these disallow colorings where two adjacent nodes get the same color. All the constraints in $\mathscr{C}$ are clearly 2-simple. We now prove that $\mathscr{C}$ is consistent if and only if G is 3-colorable.

Assume that $\mathscr{C}$ is consistent, and let f be some distribution satisfying $\mathscr{C}$. Consider the probability

$$f(\{X_{i,1} = 1, X_{i,2} = 0, X_{i,3} = 0\}) = f(\{X_{i,1} = 1, X_{i,2} = 0\}) - f(\{X_{i,1} = 1, X_{i,2} = 0, X_{i,3} = 1\}).$$

The latter probability is at most $f(\{X_{i,1} = 1, X_{i,3} = 1\})$, which by the constraints of the first type is 0. Therefore,

$$f(\{X_{i,1} = 1, X_{i,2} = 0, X_{i,3} = 0\}) = f(\{X_{i,1} = 1, X_{i,2} = 0\}) = \tfrac{1}{3}.$$

Similar reasoning allows us to conclude that $f(\{X_{i,1} = 0, X_{i,2} = 1, X_{i,3} = 0\}) = f(\{X_{i,1} = 0, X_{i,2} = 0, X_{i,3} = 1\}) = \tfrac{1}{3}$, so that $f(\{X_{i,1} = 0, X_{i,2} = 0, X_{i,3} = 0\}) = 0$. Now, pick some arbitrary point $\boldsymbol{x} \in S(f)$, and define $\gamma(v_i)$ to be b if and only if $x_{i,b} = 1$. Due to the reasoning above, there is a unique such b for every i, so that this defines a coloring of the graph. Now, consider any edge $(v_i, v_j) \in E$, and assume by contradiction that $\gamma(v_i) = \gamma(v_j) = b$. Then, $x_{i,b} = x_{j,b} = 1$, so that $f(\{X_{i,b} = 1, X_{j,b} = 1\}) \geq f(\boldsymbol{x}) > 0$, violating a constraint of the second type. Therefore, γ is a well-defined legal coloring.

Now, assume that there exists a legal coloring γ of G. Let $\pi^1, \ldots, \pi^6$ be the six permutations of $\{1, 2, 3\}$. We define f to be the uniform distribution over six points $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^6$: for $k = 1, \ldots, 6$, $i = 1, \ldots, n$, and $b = 1, 2, 3$, we define $x^k_{i,b} = 1$ if and only if $\pi^6(\gamma(v_i)) = b$. It is simple to verify, by straightforward symmetry considerations, that the resulting distribution f satisfies all the constraints in $\mathscr{C}$. □

To prove a matching upper bound, we must again make a simple assumption about the representation of the input.

DEFINITION 2.7. *An event Q is said to be* polynomially checkable *if membership of any point $\boldsymbol{x} \in \Omega$ in Q can be checked in polynomial time.*

PROPOSITION 2.8. *If all the constraints in $\mathscr{C}$ pertain to polynomially checkable events, then the consistency of $\mathscr{C}$ can be decided in nondeterministic polynomial time.*

Proof. The algorithm guesses a subset $T \subseteq \Omega$ of cardinality $|\mathscr{C}|$. It then constructs in polynomial time a system of equations corresponding to the constraints in $\mathscr{C}$ restricted to the variables in T (the other variables are set to 0). Given the initial guess, this system can be constructed in polynomial time, since, for each constraint and each point in T, it takes polynomial time to check whether the point appears in the constraint. The algorithm then attempts to find a nonnegative solution to this system. Such a solution exists if and only if there exists a manageable distribution whose sample space is (contained in) T. By Theorem 2.3, we know that a set of constraints is consistent if and only if it is satisfied by a manageable distribution, that is, a distribution over some sample space T of cardinality not greater than $|\mathscr{C}|$. Therefore, $\mathscr{C}$ is consistent if and only if one of these subsystems has a nonnegative solution. □

Since simple constraints are always polynomially checkable (using the appropriate representation), we obtain the following corollary.

COROLLARY 2.9. *For an arbitrary set $\mathscr{C}$ of simple constraints, the problem of recognizing the consistency of $\mathscr{C}$ is* NP-*complete.*

3. Independence constraints. An important special case was already discussed in the Introduction. Suppose that all the members of $\mathscr{C}$ are independence constraints arising from a known[9] fixed set of values

$$\{p_{ib}\colon i = 1, \ldots, n;\ b = 0, \ldots, r-1\},$$

where p_{ib} represents Pr $(X_i = b)$, and therefore $\sum_{b=0}^{r-1} p_{ib} = 1$ for all i, and $p_{ib} \geq 0$ for all i, b. In this case, we can construct in strongly polynomial time a manageable distribution satisfying $\mathscr{C}$. We note that the distribution we construct does not necessarily satisfy the additional constraints that Pr $(\{X_i = b\}) = p_{ib}$. If it is necessary that these constraints be satisfied, they must be put explicitly into $\mathscr{C}$.

We first define the concept of a *projected event.* Consider an event

$$Q = \{X_{i_1} = b_1, \ldots, X_{i_k} = b_k\}.$$

Let $l (1 \leq l \leq n)$ be an integer and denote by $q = q(l)$ the maximal index such that $i_q \leq l$. The *l-projection* of Q is defined as

$$\Pi_l(Q) = \{X_{i_1} = b_1, \ldots, X_{i_k} = b_q\}.$$

Intuitively, the l-projection of a constraint is its restriction to the variables $X_1, \ldots, X_l$. For example, if Q is $\{X_1 = 0, X_4 = 1, X_7 = 1\}$, then $\Pi_3(Q) = \{X_1 = 0\}$ and $\Pi_4(Q) = \{X_1 = 0, X_4 = 1\}$. Analogously, we call $I(\Pi_l(Q))$ the l-projection of the constraint $I(Q)$. Finally, for a set of independence constraints $\mathscr{C}$, $\Pi_l(\mathscr{C})$ is the set of the l-projections of the constraints in $\mathscr{C}$.

We now recursively define a sequence of distributions $f_0, f_1, \ldots, f_n$, such that, for each l $(l = 0, \ldots, n)$, the following conditions hold:

(i) f_l is a distribution on $\{0, \ldots, r-1\}^l$,

(ii) f_l satisfies $\Pi_l(\mathscr{C})$,

(iii) $|S(f_l)| \leq c$.

The distribution f_n is clearly the desired one.

We begin by defining f_0, which is a distribution on $\{0, \ldots, r-1\}^0 = \{(\ )\}$ (the singleton set containing the empty sequence). The only possible definition is $f_0((\ )) = 1$. This clearly satisfies all the requirements.

[9] The assumption that the p_{ib}'s are known is a necessary one; see Theorem 3.3.

Now assume that f_{l-1} (for $l \geq 1$) satisfies the above requirements and define an intermediate distribution g_l by

$$g_l(x_1, \ldots, x_{l-1}, b) = f_{l-1}(x_1, \ldots, x_{l-1}) \cdot p_{lb}$$

for $b = 0, \ldots, r-1$.

LEMMA 3.1. *If f_{l-1} satisfies $\Pi_{l-1}(\mathscr{C})$, then g_l satisfies $\Pi_l(\mathscr{C})$.*

Proof. We prove that g_l satisfies every constraint in $\Pi_l(\mathscr{C})$. Let $I(Q)$ be an arbitrary constraint in $\mathscr{C}$ and suppose that $Q = \{X_{i_1} = b_1, \ldots, X_{i_k} = b_k\}$. For simplicity, we denote $Q^j = \Pi_j(Q)$ ($j = 1, \ldots, n$). Let r be the maximal index such that $i_r \leq l-1$. By the assumption,

$$f_{l-1}(Q^{l-1}) = \prod_{j=1}^{r} p_{i_j b_j}.$$

We distinguish two cases.

Case I. Q mentions the variable X_l. In this case, $i_{r+1} = l$, and

$$\begin{aligned} Q^l &= \{X_{i_1} = b_1, \ldots, X_{i_r} = b_r, X_{i_{r+1}} = b_{r+1}\} \\ &= \{(x_1, \ldots, x_{l-1}, b_{r+1}) : (x_1, \ldots, x_{l-1}) \in Q^{l-1}\}. \end{aligned}$$

Therefore,

$$\begin{aligned} g_l(Q^l) &= \sum_{(x_1, \ldots, x_{l-1}) \in Q^{l-1}} g_l(x_1, \ldots, x_{l-1}, b_{r+1}) \\ &= \sum_{(x_1, \ldots, x_{l-1}) \in Q^{l-1}} f_{l-1}(x_1, \ldots, x_{l-1}) \cdot p_{lb_{r+1}} \\ &= f_{l-1}(Q^{l-1}) \cdot p_{lb_{r+1}} \\ &= \prod_{j=1}^{r} p_{i_j b_j} \cdot p_{lb_{r+1}} = \prod_{j=1}^{r+1} p_{i_j b_j}. \end{aligned}$$

Thus, g_l satisfies the constraint $I(Q^l)$.

Case II. Q does not mention the variable X_l. In this case,

$$\begin{aligned} Q^l &= \{X_{i_1} = b_1, \ldots, X_{i_r} = b_r\} \\ &= \{(x_1, \ldots, x_l) : (x_1, \ldots, x_{l-1}) \in Q^{l-1}, x_l \in \{0, \ldots, r-1\}\}. \end{aligned}$$

Therefore,

$$\begin{aligned} g_l(Q^l) &= \sum_{b \in \{0, \ldots, r-1\}} \sum_{(x_1, \ldots, x_{l-1}) \in Q^{l-1}} g_l(x_1, \ldots, x_{l-1}, b) \\ &= \sum_{b \in \{0, \ldots, r-1\}} \sum_{(x_1, \ldots, x_{l-1}) \in Q^{l-1}} f_{l-1}(x_1, \ldots, x_{l-1}) \cdot p_{lb} \\ &= \sum_{b \in \{0, \ldots, r-1\}} p_{lb} \cdot f_{l-1}(Q^{l-1}) \\ &= f_{l-1}(Q^{l-1}) = \prod_{j=1}^{r} p_{i_j b_j}. \end{aligned}$$

Again, g_l satisfies the constraint $I(Q^l)$. □

If $|S(f_{l-1})| \leq c$, then $|S(g_l)| \leq rc$, since each point with positive probability in $S(f_{l-1})$ yields at most r points with positive probabilities in $S(g_l)$. Thus, g_l satisfies requirements (i) and (ii), but may not satisfy requirement (iii). However, g_l is a nonnegative solution to the system of linear equations defined by $\Pi_l(\mathscr{C})$. Therefore, we may use Algorithm 1 or the algorithm of Beling and Megiddo [6] to reduce the cardinality of the sample space to c as described in §2. Let f_l be the resulting distribution. It clearly satisfies all three requirements. We thus obtain the following theorem.

THEOREM 3.2. *Given a set of independence constraints, we can construct a manageable distribution f satisfying $\mathscr{C}$ in strongly polynomial time using $O(rnc^{2.62})$ arithmetic operations.*

Proof. The distribution f_n constructed as above is clearly a manageable distribution satisfying $\mathscr{C}$. The construction takes n iterations. Iteration l requires at most $O(rc)$ operations to create g_l from f_{l-1}. It requires at most $O(|S(g_l)|c^{1.62}) = O(rc\cdot c^{1.62}) = O(rc^{2.62})$ arithmetic operations for running the algorithm of Beling and Megiddo to reduce g_l to f_l, as in Theorem 2.4. Therefore, the entire algorithm runs in $O(rnc^{2.62})$ arithmetic operations. The number of operations does not depend on the magnitudes of the numbers in the input. To prove that the algorithm is strongly polynomial, it remains to show that the magnitudes of the numbers used in the algorithm are polynomial in the input size. Each distribution f_l is a basic solution to the system of linear equations defined by $\Pi_l(\mathscr{C})$. The numbers used in describing this system are 1's, 0's, and products of polynomially many p_{ib}'s. Hence, their magnitudes are all polynomial in the size of the input. Since the numbers in a basic solution to a system always have polynomial length in the size of the system, we conclude that the magnitudes of the numbers in each f_l are polynomial in the size of the input. The intermediate phases—creating g_{l+1} and running the algorithm of Beling and Megiddo—do not cause blowup, since the latter is known to be strongly polynomial. □

As we mentioned, our algorithm can easily be extended to operate on random variables with ranges of different sizes. Let r_i be the number of values in the range of X_i. The sample space of g_l will consist of vectors $(x_1, \ldots, x_{l-1}, b)$, where $(x_1, \ldots, x_{l-1}) \in S(f_{l-1})$ and $b \in \{0, \ldots, r_l - 1\}$. Then $|S(g_l)| \leq r_l|\mathscr{C}|$. The proof goes through as before, but the number of operations in iteration l is $O(r_l c^{2.62})$. The total number of operations is $O((\sum_{l=1}^{n} r_l)c^{2.62}) = O(rnc^{2.62})$, where $r = \max\{r_1, \ldots, r_n\}$. The cardinality of the resulting sample space is still $|\mathscr{C}|$.

The algorithm can also deal with more general constraints with no change. In particular, it can deal with *combinatorial rectangles*, as described by Even et al. [9] and by Linial et al. [11]. A combinatorial rectangle is an independence constraint over an event of the form

$$\{[X_{i_1} \in R_1], \ldots, [X_{i_k} \in R_k]\},$$

where R_j is a subset of $\{0, \ldots, r-1\}$ (or of $\{0, \ldots, r_{i_j} - 1\}$ in the more general case). The proof remains essentially unchanged, except for minor modifications to deal with the fact that the "right" probabilities for the events (their probabilities under the assumption of independence) are different. For example, the probability of the event above would be

$$\prod_{j=1}^{k}\left(\sum_{b\in R_j} p_{i_j b}\right).$$

The complexity of the algorithm for this case and the size of the resulting sample space remain as in Theorem 3.2. Karger and Koller [11] show that this construction can be further generalized to deal with a far more general class of constraints.

Throughout this section, we have assumed that the p_{ib}'s are known. This assumption is important in view of the following theorem, which states that, if this is not the case, it is NP-hard to verify whether all of a given set of constraints are independence constraints.

THEOREM 3.3. *It is* NP-*hard to recognize whether, for a given set of* 2-*simple constraints* $\mathscr{C}$, *there exists a set* $P = \{p_{ib}\}$ *such that all the members of* $\mathscr{C}$ *are independence constraints relative to* P.

Proof. As in the proof of Proposition 2.6, we use a reduction from the problem of 3-colorability. In this proof, however, we use different variables and a different set of constraints $\mathscr{C}$. Let $G = (V, E)$ be a graph, with $V = \{v_1, \ldots, v_n\}$. We construct a set of 2-simple constraints over the 3-valued variables $X_1, \ldots, X_n$. These constraints essentially say that the probability that two neighboring vertices get the same color is 0:

$$\mathscr{C} = \{[\Pr(\{X_i = b, X_j = b\}) = 0] : (v_i, v_j) \in E; b \in \{0, 1, 2\}\}.$$

We claim that the constraints in $\mathscr{C}$ are independence constraints with respect to some P if and only if G is 3-colorable. Clearly, if the constraints in $\mathscr{C}_G$ are independence constraints relative to some P, then they are satisfiable. Let f be some distribution satisfying $\mathscr{C}$, and let $\boldsymbol{x}$ be some (arbitrary) point in $S(f)$. Define $\gamma(v_i) = x_i$. If γ is not a legal coloring, then there exists an edge $(v_i, v_j) \in E$ such that $x_i = x_j = a$. But since $f(\boldsymbol{x}) > 0$, necessarily $f(\{X_i = b, X_j = b\}) > 0$, contradicting the assumption that f satisfies $\mathscr{C}$. Assume, on the other hand, that G is 3-colorable, and let γ be an appropriate coloring. Define $p_{ib} = 1$ if $\gamma(v_i) = b$, and $p_{ib} = 0$ otherwise. We show that each constraint in $\mathscr{C}$ is an independence constraint relative to these p_{ib}'s. Each constraint in $\mathscr{C}$ is of the form $\Pr(Q^b_{(v_i,v_j)}) = 0$, for some edge $(v_i, v_j) \in E$, some $b \in \{0, 1, 2\}$, and $Q^b_{(v_i,v_j)} = \{X_i = b, X_j = b\}$. The independence constraint $I(Q^b_{(v_i,v_j)})$ relative to these p_{ib}'s is $\Pr(Q^b_{(v_i,v_j)} = p_{ib} \cdot p_{jb}$. Since γ is a legal coloring, it is impossible that both $\gamma(v_i) = b$ and $\gamma(v_j) = b$. Therefore, either $p_{ib} = 0$ or $p_{jb} = 0$ and their product is necessarily 0, resulting in the desired constraint. □

This theorem can be interpreted as showing the NP-hardness of deciding whether a set of constraints is satisfied by an independent distribution. It shows that the problem is hard for 2-simple constraints over 3-valued random variables. It is also possible to prove, using a reduction in 3-SAT, that this problem is hard for 3-simple constraints over binary-valued random variables. But, unlike the problem of deciding the satisfiability of a set of constraints by an *arbitrary* distribution (Proposition 2.6), the problem is *not* NP-hard for the case of 2-simple constraints over binary-valued random variables. In this case, a numeric variant of the standard algorithm for 2-SAT can be used to solve the problem in polynomial time.

It is not clear that the problem of Theorem 3.3 is even in NP. The set P relative to which a given $\mathscr{C}$ is a set of independence constraints might contain irrational numbers even if all the numbers in the input are rational.

Example 3.4. Consider the problem of constructing a distribution over the binary-valued variables X_1, X_2, and X_3 satisfying

$$\Pr(\{X_1 = 1, X_2 = 1\}) = \tfrac{1}{2},$$

$$\Pr(\{X_1 = 1, X_3 = 1\}) = \tfrac{1}{2},$$

$$\Pr(\{X_2 = 1, X_3 = 1\}) = \tfrac{1}{2}.$$

These are independence constraints only with respect to $p_{11} = p_{21} = p_{31} = 1/\sqrt{2}$. □

In most practical cases, however, the p_{ib}'s are part of the specification of the algorithm. Thus, it is usually reasonable to assume that they are known.

4. Derandomizing algorithms. In this section, we demonstrate how the technique of §3 can be used to derandomize algorithms. We present three progressively improving ways in which the technique can be applied. For simplicity and ease of comparison, we base our analysis on a single problem: finding large independent sets in sparse hypergraphs.

The problem description and the randomized algorithm for its solution are taken from Alon, Babai, and Itai [3]. Note that a deterministic polynomial-time algorithm for this problem is known [2].

A *d-uniform hypergraph* is a pair $\mathscr{H} = (V, \mathscr{E})$, where $V = \{v_1, \ldots, v_n\}$ is a set of *vertices* and $\mathscr{E} = \{E_1, \ldots, E_m\}$ is a collection of subsets of V, each of cardinality d, which are called *edges*. (For simplicity, we restrict attention to d-uniform hypergraphs; a similar analysis goes through in the general case.) A subset $U \subseteq V$ is said to be independent if it contains no edge.

Algorithm 2. Independent sets in hypergraphs.

```
1. Construct a random subset $R$ of $V$.
   For each vertex $v_i \in V$:
      put $v_i$ in $R$ with probability $p = 3k/n$.
2. Modify $R$ into an independent set $U$.
   For each edge $E_j \in \mathscr{E}$ such that $E_j \subseteq R$:
      remove from $R$ some arbitrary vertex $v_i \in E_j$.
```

Consider the randomized algorithm above (k is defined later). The following theorem, due to Alon, Babai, and Itai [3], states that this algorithm finds "large" independent sets in hypergraphs with "high" probability. We only sketch the proof of this theorem, concentrating on the part that is relevant to this discussion: the constraints on the distribution assumed by the proof.

PROPOSITION 4.1 (Alon, Babai, and Itai [3]). *If $\mathscr{H} = (V, \mathscr{E})$ is a d-uniform hypergraph with n vertices and m edges, then, for $k = (1/18)(n^d/m)^{1/(d-1)}$, Algorithm 2 finds an independent set of cardinality exceeding k with probability greater than $\frac{1}{2} - 3/k$.*

Proof. For each vertex $v_i \in V$, *let* X_i be the random variable that equals 1 if $v_i \in R$ and 0 otherwise. For each edge $E_j \in \mathscr{E}$, let Y_j be the random variable that equals 1 if $E_j \subseteq R$ and 0 otherwise. The cardinality of R is $|R| = \sum_{i=1}^n X_i = X$, so $E(X) = np = 3k$.

- *If the X_i's are pairwise independent*, then the variance of X is

$$\sigma^2(X) = \sum_{i=1}^{n} \sigma^2(X_i) = np(1-p) < np = 3k. \tag{1}$$

  Thus, using Chebychev's inequality,

$$\Pr(X \leq 2k) \leq \frac{\sigma^2(X)}{k^2} < \frac{3}{k}.$$

- *If the X_i's are d-wise independent*, then, for every $j = 1, \ldots, m$,

$$E(Y_j) = \Pr\left(\bigcap_{i \in E_j} \{X_i = 1\}\right) = p^d. \tag{2}$$

  Let $Y = \sum_{j=1}^m Y_j$ denote the number of edges contained in R. Computation shows that $\Pr(Y \geq k) < \frac{1}{2}$.

If R contains at least $2k$ vertices after the first stage in the algorithm and at most k vertices are removed in the second stage, then the independent set constructed by the algorithm has cardinality at least k. This has probability at least

$$\Pr(\{Y < k\} \cap \{X \geq 2k\}) > \frac{1}{2} - \frac{3}{k},$$

as desired. □

Derandomization I. The derandomization procedure of Alon, Babai, and Itai [3] is based on constructing a joint distribution of d-wise independent variables X_i that approximates the joint d-wise independent distribution for which $\Pr(X_i = 1) = 3k/n$ $(i = 1, \ldots, n)$. It is then necessary to analyze this approximate distribution to verify that the above correctness proof continues to hold. Our technique provides exactly the required distribution, so that no further analysis is needed. As we explained in the Introduction, this can be done by considering the set $\mathscr{C}^{\mathrm{I}}$ of the constraints[10]

$$I(\{X_{i_1} = b_1, \ldots, X_{i_d} = b_d\}) \colon i_1, \ldots, i_d \in \{1, \ldots, n\}, b_1, \ldots, b_d \in \{0, 1\}\}.$$

The number of these constraints is $|\mathscr{C}^{\mathrm{I}}| = \binom{n}{d}2^d = O((2n)^d)$. For fixed d, this number is polynomial in n, resulting in a sample space of polynomial size (in fact, the size of the sample space is comparable to that achieved in [3]). Therefore, the algorithm runs in polynomial time, including both the phase of constructing the sample space and the phase of running step 2 of Algorithm 2 on each point of this space until a sufficiently large independent set is found.

Derandomization II. A closer examination of the proof reveals that not all the $\binom{n}{d}$ neighborhoods of cardinality d must be independent. For (2) to hold, it suffices that only the X_i's associated with vertices in the same edge be independent. If $E_j = \{v_{i_1}, \ldots, v_{i_d}\}$, let $\mathscr{C}_j$ denote the set of 2^d independence constraints

$$\{I(\{X_{i_1} = b_1, \ldots, X_{i_d} = b_d\}) \colon b_1, \ldots, b_d \in \{0, 1\}\}.$$

On the other hand, for (1) to hold, the X_i's must still be pairwise independent. Let $\mathscr{C}^2$ denote the set of $4\binom{n}{2}$ constraints

$$\{I(\{X_{i_1} = b_1, X_{i_2} = b_2\}) \colon i_1, i_2 \in \{1, \ldots, n\}, b_1, b_2 \in \{0, 1\}\}.$$

Thus, the following set of constraints suffices:

$$\mathscr{C}^{\mathrm{II}} = \mathscr{C}^2 \cup \bigcup_{E_j \in \mathscr{E}} \mathscr{C}_j.$$

More precisely, if the set $\mathscr{C}^{\mathrm{II}}$ is satisfied, then the proof of Proposition 4.1 goes through, and the resulting sample space must contain a point that is good for this hypergraph. Since the number of constraints is

$$|\mathscr{C}^{\mathrm{II}}| = |\mathscr{C}^2| + \sum_{E_j \in \mathscr{E}} |\mathscr{C}_j| = 4\binom{n}{2} + m2^d,$$

this results in a polynomial-time algorithm for $d = O(\log n)$. This algorithm therefore applies to a larger class of graphs than that presented by Alon, Babai, and Itai [3]. At first, it seems that, as we have polynomially many neighborhoods of logarithmic size, Schulman's technique [16] can also be used in this case. However, his approach is limited to (uniformly distributed) random bits, so it does not apply to this algorithm. The results of Berger and Rompel [7] and of Motwani, Naor, and Naor [13], however, provide a polynomial-time algorithm for $d = O(\text{polylog } n)$. Their results use a completely different technique and cannot be extended to handle larger values of d.

Derandomization III. A yet closer examination of the proof of Proposition 4.1 reveals that (2) does not require complete independence of the neighborhood associated with

[10] Theoretically, we also need to include the constraints $\Pr(\Omega) = 1$ and $\Pr(\{X_i = 1\}) = p$ for all i. However, these are implied by the other constraints in $\mathscr{C}^{\mathrm{I}}$. This will also be the case for the later sets of constraints $\mathscr{C}^{\mathrm{II}}$ and $\mathscr{C}^{\mathrm{III}}$.

the edge E_j. It suffices to constrain the probability of the event "all the vertices in E_j are in R" (the event corresponding to the random variable Y_j in the proof). That is, for $E_j = \{v_{i_1}, \ldots, v_{i_d}\}$, we need only the independence constraint over the event

$$Q_j = \{X_{i_1} = 1, \ldots, X_{i_d} = 1\}.$$

This is a simple event that defines an independence constraint of the type to which our technique applies. We conclude that the following set of constraints suffices for the analysis of Proposition 4.1 to go through:

$$\mathscr{C}^{\text{III}} = \mathscr{C}^2 \cup \{I(Q_j) : E_j \in \mathscr{E}\}.$$

The number of constraints

$$|\mathscr{C}^{\text{III}}| = 4\binom{n}{2} + m$$

is polynomial in n and m, regardless of d. Therefore, this results in a deterministic polynomial-time algorithm for finding large independent sets in arbitrary uniform hypergraphs.

5. Conclusions and open questions. We have presented a new approach to constructing distributions with small sample spaces. Our technique constructs a distribution tailored precisely to the required constraints. The construction is based on an explicit representation of the constraints as a set of linear equations over the distribution. It enables us to construct sample spaces for arbitrary distributions over discrete random variables, which are precise (not approximations) and sometimes considerably smaller than sample spaces constructed using previously known techniques. This construction can be done in polynomial time for a large class of practical problems—those problems that can be described using only independence constraints.

A number of open questions arise immediately from our results.

- Schulman's approach constructs a sample space whose size depends not on the total number of neighborhoods involved in constraints, but on the maximum number of such neighborhoods in which a particular variable appears. Perhaps the size of the sample space in our approach can similarly be reduced to depend on the maximum number of independence constraints in which a variable X_i participates.
- We mentioned in the Introduction that the nature of our approach generally prevents a precomputation of the manageable distribution. However, our approach shows the existence of manageable distributions that are useful in general contexts. For example, for every n, d, and p, we show the existence of a d-wise independent distribution over n binary random variables such that $\Pr(X_i = 1) = p$ for all i. It would be useful to come up with an explicit construction for this class of distributions.
- Our technique constructs distributions that precisely satisfy a given set of arbitrary independence constraints. It is natural to ask if our results can be improved by only requiring the distribution to approximately satisfy these constraints. In particular, it may be possible to construct approximate distributions faster, or in parallel (see [11]), or over smaller sample spaces. We note that the original d-wise independent constructions [3], [10], [12] can be viewed as precisely satisfying the d-wise independence constraints but approximately satisfying the constraints on $\Pr(X_i = b)$. In contrast, the nearly-independent constructions [4], [5], [9], [14] can be viewed as approximately satisfying the d-wise independence constraints. Thus, they all provide an answer to this question for certain types of constraint-sets $\mathscr{C}$ and certain restrictions on which constraints can be approximated.
- Combined with our inability to precompute the distribution, the sequential nature of our construction prevents its use for derandomization of parallel algorithms. Paral-

lelizing the construction could open up many application areas for this approach (see [11]).

Acknowledgments. The authors thank Joe Kilian for suggesting a considerably simplified proof for Proposition 2.6. We also thank Yossi Azar and David Karger for stimulating discussions, and Howard Karloff, Moni Naor, and Sundar Vishwanathan for useful comments on previous versions of this paper.

REFERENCES

[1] L. ADLEMAN, *Two theorems on random polynomial time*, in Proc. of the 19th Annual Symposium on Foundations of Computer Science, 1978, pp. 75–83.

[2] N. ALON, private communication, 1992.

[3] N. ALON, L. BABAI, AND A. ITAI, *A fast and simple randomized parallel algorithm for the maximal independent set problem*, J. Algorithms, 7 (1986), pp. 567–583.

[4] N. ALON, O. GOLDREICH, J. HASTAD, AND R. PERALTA, *Simple constructions of almost k-wise independent random variables*, in Proc. of the 31st Annual Symposium on Foundations of Computer Science, 1990, pp. 544–553.

[5] Y. AZAR, R. MOTWANI, AND J. NAOR, *Approximating arbitrary probability distributions using small sample spaces*, unpublished manuscript, 1992.

[6] P. A. BELING AND N. MEGIDDO, *Using fast matrix multiplication to find basic solutions*, Tech. Report RJ 9234, IBM Research Division, 1993.

[7] B. BERGER AND J. ROMPEL, *Simulating* ($\log^c n$)*-wise independence in* NC, in Proc. of the 30th Annual Symposium on Foundations of Computer Science, 1989, pp. 2–7.

[8] B. CHOR, O. GOLDREICH, J. HASTAD, J. FRIEDMAN, S. RUDICH, AND R. SMOLENSKY, *t-resilient functions*, in Proc. of the 26th Annual Symposium on Foundations of Computer Science, 1985, pp. 396–407.

[9] G. EVEN, O. GOLDREICH, M. LUBY, N. NISAN, AND B. VELIČKOVIĆ, *Approximations of general independent distributions*, in Proceedings of the 24th Annual ACM Symposium on Theory of Computing, 1992, pp. 10–16.

[10] A. JOFFE, *On a set of almost deterministic k-independent random variables*, Ann. Probab., 2 (1974), pp. 161–162.

[11] D. R. Karger and D. Koller, *A (de)randomized derandomization technique*, 1994, in preparation.

[12] N. LINIAL, M. LUBY, M. SAKS, AND D. ZUCKERMAN, *Efficient construction of a small hitting set for combinatorial rectangles in high dimension*, in Proc. of the 25th Annual ACM Symposium on Theory of Computing, 1993, to appear.

[13] M. LUBY, *A simple parallel algorithm for the maximal independent set problem*, SIAM J. Comput., 15 (1986), pp. 1036–1053.

[14] R. MOTWANI, J. NAOR, AND M. NAOR, *The probabilistic method yields deterministic parallel algorithms*, in Proc. of the 30th Annual Symposium on Foundations of Computer Science, 1989, pp. 8–13.

[15] J. NAOR AND M. NAOR, *Small-bias probability spaces: Efficient constructions and applications*, in Proc. of the 22nd Annual ACM Symposium on Theory of Computing, 1990, pp. 213–223; J. Comput. System Sci., to appear.

[16] P. RAGHAVAN, *Probabilistic construction of deterministic algorithms: Approximating packing integer problems*, J. Comput. System Sci., 37 (1988), pp. 130–143.

[17] L. J. SCHULMAN, *Sample spaces uniform on neighborhoods*, in Proc. of the 24th Annual ACM Symposium on Theory of Computing, 1992, pp. 17–25.

[18] J. SPENCER, *Ten Lectures on the Probabilistic Method*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 275–283, May 1994

013

THE MINIMUM SATISFIABILITY PROBLEM*

RAJEEV KOHLI[†], RAMESH KRISHNAMURTI[‡], AND PRAKASH MIRCHANDANI[§]

Abstract. This paper shows that a minimization version of satisfiability is strongly NP-hard, even if each clause contains no more than two literals and/or each clause contains at most one unnegated variable. The worst-case and average-case performances of greedy and probabilistic greedy heuristics for the problem are examined, and tight upper bounds on the performance ratio in each case are developed.

Key words. minimum satisfiability, satisfiability, heuristics, probabilistic algorithms, average performance, Horn formulae

AMS subject classification. 68Q25

1. Introduction. The satisfiability problem is perhaps one of the most well-studied problems in logic theory. Given a set V of Boolean (true/false) variables and a collection D of clauses over V, the satisfiability problem is to determine if there is a truth assignment that satisfies all clauses in D. The problem is NP-complete even when every clause in D has at most three literals (Even, Itai, and Shamir [2]). The maximum satisfiability (MAXSAT) problem is an optimization version of satisfiability that seeks a truth assignment to maximize the number of satisfied clauses (Johnson [5]). The MAXSAT problem is NP-hard even when every clause contains at most two literals (Garey, Johnson, and Stockmeyer [4]).

In this paper, we consider the following complement of the MAXSAT problem. Given a set U of Boolean variables and a collection C of clauses over U, find a truth assignment that minimizes the number of satisfied clauses. We call this the minimum satisfiability (MINSAT) problem. The existence of a truth assignment for the MINSAT problem that satisfies no clause can be trivially determined because such an assignment exists only if each variable or its negation appears in no clause. Similarly, if each clause contains one literal, the solution to the MINSAT problem is readily obtained by setting a variable true if it occurs in less clauses than its negation and setting the variable false otherwise. However, we show that, in general, the MINSAT problem is NP-hard, even if every clause contains no more than two literals. We then consider two heuristics for solving the problem. The first is a greedy heuristic similar to a procedure described by Johnson [5] for the MAXSAT problem. The second is a probabilistic greedy heuristic similar to an algorithm by Kohli and Krishnamurti [6] for the MAXSAT problem. Like the greedy heuristic, the probabilistic greedy heuristic selects a truth assignment one variable at a time. Unlike the greedy heuristic, the probabilistic greedy heuristic introduces a chance element in selecting a truth assignment, forcing a trade-off between the value of a nonoptimal solution and the probability of its selection. We characterize the worst-case and average-case performances of the two heuristics and show that, while the probabilistic greedy heuristic can select an arbitrarily bad assignment in the worst case, on average it satisfies no more than twice the optimal number of clauses, regardless of the data-generating distribution. On the other hand, if each clause contains at most s literals, the greedy heuristic satisfies no more than s times the number of clauses satisfied by the optimal assignment. However, the average performance of the greedy heuristic depends

* Received by the editors August 13, 1991; accepted for publication (in revised form) May 11, 1993.

[†] Graduate School of Business, Columbia University, 506 Uris Hall, New York, New York 10027.

[‡] School of Computing Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. This author was supported by Natural Sciences and Engineering Research Council of Canada grant OGP36809.

[§] Joseph M. Katz School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

upon the minimum probability with which it selects an optimal assignment at any step. As this probability decreases (increases), the average performance of the greedy heuristic tends to the worst-case (optimal) solution.

In §2 we show that the MINSAT problem is NP-hard. In §3 we formally describe the greedy heuristic and analyze its worst-case and average-case performances. In §4 we analyze the average-case performance of the probabilistic greedy heuristic. We derive tight upper bounds on the performance ratio in each case. In §5 we extend our results to Horn formulae, in which each clause has no more than one unnegated variable.

2. Complexity. We show in Theorem 1 that the following decision problem, called the 2-MINSAT problem, in which each clause contains exactly two literals, is NP-complete. It follows that the MINSAT problem is NP-hard if each clause contains at least two literals.

2-MINSAT.

Instance. Set U of k variables, collection C of n clauses over U such that each clause $c \in C$ has $|c| = 2$ literals, positive integer $n^* \le n$.

Question. Is there a truth assignment for U that satisfies no more than n^* clauses in C?

In Theorem 1, we transform the following 2-MAXSAT problem, which is NP-complete (Garey and Johnson [3, pp. 259–260]), to the 2-MINSAT problem.

2-MAXSAT.

Instance. Set V of h variables, collection D of l clauses over V such that each clause $d \in D$ has $|d| = 2$ literals, positive integer $l^* \le l$.

Question. Is there a truth assignment for V that satisfies at least l^* clauses in D?

THEOREM 1. *The* 2-MINSAT *problem is* NP-*complete.*

Proof. Given a *yes* instance of the 2-MINSAT problem, we can simply count the number of satisfied clauses and verify a *yes* instance in polynomial time. Hence the 2-MINSAT problem is in NP. To show that it is NP-complete, we transform the 2-MAXSAT problem to the 2-MINSAT problem as follows.

Let $d = q_a \vee q_b$ be a clause in an instance of the 2-MAXSAT problem, where q_a and q_b denote either variables in V or their negations. For each clause $d \in D$, define a variable w_d. Let $W = \{w_d | d \in D\}$. For each clause $d = q_a \vee q_b$ of 2-MAXSAT, define a pair of clauses $c_{1d}, c_{2d} \in C$ for 2-MINSAT, where

$$c_{1d} = \bar{q}_a \vee w_d, \qquad c_{2d} = \bar{q}_b \vee \bar{w}_d.$$

Let $C = \{c_{1d}, c_{2d} | d \in D\}$. *Let* $n^* = 2l - l^*$. Thus, given an instance of the 2-MAXSAT problem defined over the set V of h variables and the set D of l clauses, we construct in polynomial time an instance of the 2-MINSAT problem defined over the set $U = V \cup W$ of $k = h + l$ variables and the set C of $n = 2l$ clauses. We now show that no less than l^* clauses can be satisfied by a truth assignment for 2-MAXSAT if and only if no more than n^* clauses can be satisfied by a truth assignment for 2-MINSAT.

Suppose there exists a truth assignment for 2-MAXSAT that satisfies $m \ge l^*$ clauses. Clause $d = q_a \vee q_b$ is not satisfied by an assignment if and only if both q_a and q_b are false, in which case, both c_{1d} and c_{2d} are satisfied. On the other hand, clause d is satisfied if and only if at least one of q_a or q_b is true. If both q_a and q_b are true, then any truth assignment for w_d satisfies exactly one of c_{1d} and c_{2d}. If q_a is true and q_b is false, then a false assignment for w_d satisfies c_{2d} but not c_{1d}. Similarly, if q_a is false and q_b is true, then a true assignment for w_d satisfies c_{1d} but not c_{2d}. Thus, if clause d is satisfied, a suitable truth assignment for w_d ensures that only one of c_{1d} or c_{2d} is satisfied. Consequently, if a truth assignment for 2-MAXSAT satisfies exactly $m \ge l^*$ clauses, then a suitable truth assignment for each $w_d \in W$ ensures that exactly $m + 2(l - m) = 2l - m \leq 2l - l^* = n^*$ clauses are satisfied in the corresponding instance of 2-MINSAT. Hence every *yes* instance of the 2-MAXSAT problem corresponds to a *yes* instance of the 2-MINSAT problem.

Now suppose that a truth assignment for the 2-MINSAT problem satisfies no more than $n^* = 2l - l^*$ clauses. Consider the pair of clauses $c_{1d} = \bar{q}_a \vee w_d$ and $c_{2d} = \bar{q}_b \vee \bar{w}_d$. Any assignment of w_d ensures that at least one of these two clauses is satisfied. Let y denote the number of pairs $c_{1d}, c_{2d}, d \in D$, where exactly one clause in the pair is satisfied by a given assignment for the 2-MINSAT problem. Without loss of generality, let $c_{1d} = \bar{q}_a \vee w_d$ be the clause that is not satisfied. Then q_a must be true, which implies that clause $d = q_a \vee q_b$ must be satisfied. Hence, at least y clauses in D must be satisfied. Since there are l pairs c_{1d}, c_{2d}, there are $l - y$ such pairs in which both clauses are satisfied. Thus, the number of clauses in C that are satisfied is $2(l - y) + y \leq n^* = 2l - l^*$, which implies that $y \geq l^*$. Thus, if $n^* = 2l - l^*$ clauses in C are satisfied for the 2-MINSAT problem, then at least l^* clauses in D are satisfied for the 2-MAXSAT problem. Hence, every *yes* instance of the 2-MINSAT problem corresponds to a *yes* instance of the 2-MAXSAT problem. □

3. The greedy heuristic. Let $u_1, u_2, \ldots, u_k$ denote an arbitrary ordering of the k variables in U for the MINSAT problem. Given any ordering of the variables, the greedy heuristic sequentially selects an assignment for each variable to satisfy the smallest number of additional clauses. We begin by describing the greedy heuristic more formally below.

Initialization (*Step* 1). Let $C_1 = C$ denote the set of all clauses in an instance of the MINSAT problem. Let $C_1(u_1)$ denote the subset of clauses in C_1 that contain variable u_1. Let $C_1(\bar{u}_1)$ denote the subset of clauses in C_1 that contain variable $\bar{u}_1$. Let x_1 and y_1 denote the number of clauses in sets $C_1(u_1)$ and $C_1(\bar{u}_1)$. At the first step, the greedy heuristic selects the partial assignment u_1 (i.e., assigns u_1 to be true) if $x_1 < y_1$. Otherwise, it selects the partial assignment $\bar{u}_1$ (i.e., assigns u_1 to be false).[1] All clauses satisfied by the partial assignment are eliminated. Let C_2 denote the set of clauses not satisfied at the end of Step 1. Thus,

$$C_2 = \begin{cases} C_1 \backslash C_1(u_1) & \text{if } u_1 \text{ is selected at Step 1,} \\ C_1 \backslash C_1(\bar{u}_1) & \text{if } \bar{u}_1 \text{ is selected at Step 1.} \end{cases}$$

Recursion (*Step j*). Let C_j denote the set of clauses that are not satisfied at the end of Step $j - 1$. Let $C_j(u_j)$ denote the subset of clauses in C_j that contain u_j. Let $C_j(\bar{u}_j)$ denote the subset of clauses in C_j that contain $\bar{u}_j$. Let x_j and y_j denote the number of clauses in sets $C_j(u_j)$ and $C_j(\bar{u}_j)$. At Step j, the greedy heuristic includes u_j in the partial assignment (i.e., assigns u_j to be true) if $x_j < y_j$. Otherwise, it includes $\bar{u}_j$ in the partial assignment (i.e., assigns u_j to be false). All clauses satisfied by the partial assignment are eliminated. Let C_{j+1} denote the set of clauses not satisfied at the end of Step j. Thus,

$$C_{j+1} = \begin{cases} C_j \backslash C_j(u_j) & \text{if } u_j \text{ is selected at Step } j, \\ C_j \backslash C_j(\bar{u}_j) & \text{if } \bar{u}_j \text{ is selected at Step } j. \end{cases}$$

Termination Step. Stop if $C_{j+1} = \phi$ or if $j = k$.

Let $|c_i|$ denote the number of literals in clause c_i. Let $s = \max_i |c_i|$ denote the maximum number of literals in any clause. Let r denote the performance ratio for the greedy heuristic, i.e., the ratio of the number of clauses satisfied by the assignment selected

[1] Note that we assign u_j to be false if $x_j = y_j$. This simplifies the subsequent worst-case analysis of the heuristic, where we assume, without loss of generality, that the optimal solution is to set each variable true.

by the greedy heuristic to the number of clauses satisfied by an optimal assignment. Theorem 2 shows that the value of the performance ratio r is bounded from above by s. As there are no more than k literals, $s \leq k$, and it follows trivially from Theorem 2 that $r \leq k$.

THEOREM 2. *$r \leq s$ for the greedy heuristic.*

Proof. Without loss of generality, let each variable $u_j \in U$ be true in the optimal assignment. Let $C_j^* \subseteq C$ denote the subset of clauses in C that are satisfied if variable u_j is true. Let $C^* = \bigcup_{j=1}^{k} C_j^*$ denote the subset of clauses satisfied by the optimal assignment. Let m_j denote the number of clauses in the set C_j^*. Let m denote the number of clauses in set C^*. At Step j, the greedy heuristic satisfies $z_j = \min\{x_j, y_j\}$ clauses in set C_j. Hence the total number of clauses satisfied by the greedy heuristic is $\sum_{j=1}^{k} z_j$. Now $m_j \geq x_j$, which implies that

$$\sum_{j=1}^{k} z_j \leq \sum_{j=1}^{k} x_j \leq \sum_{j=1}^{k} m_j.$$

Also,

$$\sum_{j=1}^{k} m_j = \sum_{c_i \in C^*} |c_i| \leq ms.$$

Thus, an upper bound on the performance ratio of the greedy heuristic is

$$r = \frac{\sum_{j=1}^{k} z_j}{m} \leq \frac{ms}{m} = s. \qquad \square$$

To prove that the bound in Theorem 2 is tight, consider the following problem instance in which there are s variables and $s + 1$ clauses:

$$c_1 = u_1 \vee u_2 \vee \cdots \vee u_s,$$
$$c_i = \bar{u}_{i-1} \quad \text{for } 2 \leq i \leq s + 1.$$

Each variable u_j, $1 \leq j \leq s$ is true in the optimal assignment, which satisfies only one clause c_1. The greedy heuristic sets each variable u_j false, $1 \leq j \leq s$ and satisfies s clauses, $c_2, c_3, \ldots, c_{s+1}$. Hence, $r = s$. Note that, in this worst-case example, a reordering of the variables has no effect on the performance of the greedy heuristic. This example also suffices to show that, given the solution selected by the greedy heuristic, an interchange heuristic that seeks to maximally improve the solution value by replacing a literal by its negation does no better than the greedy heuristic alone.

In [5] Johnson suggests weighted greedy heuristics for the MAXSAT problem, the simplest of which ensures a worst-case error of $1/2^s$ when each clause has no less than s literals. An analogous weighted greedy heuristic for the MINSAT problem is as follows. At Step j, the greedy heuristic assigns

$$u_j \text{ true} \quad \text{if } \sum_{c_i \in C_j(u_j)} |c_i| > \sum_{c_i \in C_j(\bar{u}_j)} |c_i|,$$
$$u_j \text{ false} \quad \text{otherwise.}$$

The intuition behind the weighting is that a literal should be selected if the unsatisfied clauses that contain its negation are less likely to be satisfied by subsequent assignments to variables. However, this weighting scheme does not improve the worst-case performance of the greedy heuristic for the MINSAT problem. To illustrate, consider the following

$s + 1$ clauses defined over $2s - 1$ variables u_j, $1 \leq j \leq 2s - 1$:

$$c_1 = u_1 \vee u_2 \vee \cdots \vee u_s,$$
$$c_i = \bar{u}_{i-1} \vee \bar{u}_{s+1} \vee \bar{u}_{s+2} \vee \cdots \vee \bar{u}_{2s-1}, \qquad 2 \leq i \leq s + 1.$$

Each variable u_j, $1 \leq j \leq 2s - 1$ is true in the optimal assignment, which satisfies only one clause, c_1. At each step j of the greedy heuristic, an equal number of clauses are satisfied by setting u_j true or false. Regardless of the truth assignment for variable u_j at Step j, there are an equal number of literals in each remaining (unsatisfied) clause. Consequently, any weighting of the clauses does not affect the greedy solution, which in the worst case assigns a "false" value to each variable u_j, $1 \leq j \leq 2s - 1$ and satisfies the s clauses $c_2, c_3, \ldots, c_{s+1}$. Note, however, that the ordering of the variables is critical to this worst-case example. Thus, the performance ratio for this weighted greedy heuristic for the MINSAT problem is no better than s. Indeed, other weighting schemes (e.g., a scheme similar to Johnson's [5] exponential weighting of clauses) also do not improve the worst-case performance of the greedy heuristic.

We next examine the average performance of the greedy heuristic. Let P_k denote the MINSAT problem defined over the set U of k variables. Let P_{k-j+1} denote the MINSAT problem at Step j of the greedy heuristic, where problem P_{k-j+1} is defined over the subset of unassigned variables $u_j, u_{j+1}, \ldots, u_k$. Following Kohli and Krishnamurti [6], we assume that there is a probability p_j with which the truth assignment selected for a variable at step j of the greedy heuristic appears in an optimal truth assignment for problem P_{k-j+1}. We assume that the p_j are independent across the steps of the greedy heuristic. However, we do not assume either the independence of the literals across clauses, or any specific data-generating distribution.

Let r_j denote the performance ratio of the greedy heuristic for problem P_j. Let $E(r_j)$ denote the expected value of r_j. Theorem 3 characterizes the lower bound on $E(r_k)$ as a function of k and $p = \min_{1 \leq j \leq k} p_j$.

THEOREM 3. *$E(r_k) \leq 1 - (1 - p)^k/p$ for the greedy heuristic.*

Proof. Without loss of generality, we assume that the optimal solution is to set variable u_j true for all j, $1 \leq j \leq k$. We prove the theorem by induction on k, the number of variables.

For $k = 1$, the greedy heuristic chooses the optimal assignment and sets u_1 true. Hence, $p = p_1 = 1$ and $E(r_k) = 1$.

Let $l \geq 1$ be an integer such that

$$E(r_k) \leq \frac{1 - (1 - p)^k}{p} \quad \text{for } k = l.$$

We show below that

$$E(r_k) \leq \frac{1 - (1 - p)^k}{p}$$

for $k = l + 1$.

Let $z_1 = \min \{x_1, y_1\}$. If the greedy heuristic sets u_1 true at Step 1, the value of the optimal solution to problem P_{k-1} is $m - z_1$, where m is the value of the optimal solution to problem P_k. However, if the greedy heuristic sets u_1 false at Step 1, the value of the optimal solution to problem P_{k-1} is bounded from above by m. In either case, after Step

1, the greedy heuristic solves an l variable MINSAT problem. Thus,

$$E(r_{l+1}) \le \frac{1}{m}(z_1 + p_1 E(r_l)(m - z_1) + (1 - p_1)E(r_l)m)$$

$$= \frac{z_1}{m}(1 - p_1 E(r_l)) + E(r_l) \le 1 + E(r_l)(1 - p_1).$$

Let $p^* = \min_{i \ge 2} p_i$. By the induction hypothesis,

$$E(r_l) \le \frac{1 - (1 - p^*)^l}{p^*}.$$

Thus,

$$E(r_{l+1}) \le 1 + \left(\frac{1 - (1 - p^*)^l}{p^*}\right)(1 - p_1).$$

Let $p = \min \{p^*, p_1\}$. Then

$$\frac{1 - (1 - p^*)^l}{p^*} \le \frac{1 - (1 - p)^l}{p} \quad \text{and} \quad 1 - p_1 \le 1 - p,$$

which implies that

$$E(r_{l+1}) \le 1 + \left(\frac{1 - (1 - p)^l}{p}\right)(1 - p) = \frac{1 - (1 - p)^{l+1}}{p}. \qquad \square$$

Note that the bound derived in Theorem 3 approaches 1 as p approaches 1 and that it approaches k as p approaches zero. As k tends to infinity, the bound on the average performance ratio for the greedy heuristic approaches $1/p$.

To prove that the bound derived in Theorem 3 is tight, consider the following example with k variables and $n = (k + 1)N + k$ clauses. The first $(k + 1)N$ clauses are

$$c_i = \begin{cases} u_1 \vee u_2 \vee \cdots \vee u_k & \text{for } 1 \le i \le N \\ \bar{u}_1 & \text{for } N + 1 \le i \le 2N, \\ \bar{u}_2 & \text{for } 2N + 1 \le i \le 3N, \\ \vdots & \\ \bar{u}_k & \text{for } kN + 1 \le i \le (k + 1)N. \end{cases}$$

The remaining k clauses are probabilistically generated, each clause containing exactly one of the k distinct variables in negated form with probability p and in unnegated form with probability $1 - p$. Specifically, clause j, $1 \le j \le k$ is given by

$$c_{(k+1)N+j} = \begin{cases} \bar{u}_j & \text{with probability } p, \\ u_j & \text{with probability } 1 - p. \end{cases}$$

Each variable u_j, $1 \le j \le k$ is true in the optimal assignment. For $N \gg k$, the expected performance ratio of the greedy heuristic can be verified to approach from below the value

$$p + 2(1 - p)p + 3(1 - p)^2 p + \cdots + (k - 1)(1 - p)^{k-2} p + k(1 - p)^{k-1} = \frac{1 - (1 - p)^k}{p}.$$

Observe that the bound on the average performance of the greedy heuristic depends upon the value of p, which can vary, depending upon the data-generating distribution. If, as in the above example, the value of p can be made close to zero, the average performance of the greedy heuristic can be made to approach its deterministic worst-case bound. In the next section, we examine a probabilistic greedy heuristic that, regardless of the data-generating distribution, never satisfies more than twice the number of clauses satisfied by an optimal assignment for the MINSAT problem.

4. Probabilistic greedy heuristic. The proposed probabilistic greedy heuristic differs from the preceding greedy heuristic by the introduction of a probabilistic element in the choice of a truth assignment for each variable. In particular, at Step j, the probabilistic greedy heuristic sets u_j to be true with probability $q_j = y_j/(x_j + y_j)$ and sets u_j to be false with probability $1 - q_j$. Thus, the probability of setting u_j true increases as x_j/y_j decreases and is 1 only if $x_j = 0$, i.e., if no additional clauses are satisfied by setting u_j true at Step j of the heuristic. However, if the number of additional clauses satisfied is greater when u_j is true than when it is false, then u_j is set true with a smaller probability than it is set false.

The following theorem shows that, on average, the number of clauses satisfied by the probabilistic greedy heuristic is no larger than twice the number of clauses satisfied by the optimal assignment.

THEOREM 4. *$E(r_k) \leq 2$ for the probabilistic greedy heuristic.*

Proof. Without loss of generality, assume that variable u_1 is true in an optimal assignment. We prove the theorem by induction on the number of variables k.

For $k = 1$, the greedy heuristic sets u_1 true with probability $q_1 = y_1/(x_1 + y_1)$ and sets u_1 false with probability $1 - q_1$. The expected number of satisfied clauses is

$$q_1x_1 + (1 - q_1)y_1 = \frac{y_1}{x_1 + y_1}x_1 + \frac{x_1}{x_1 + y_1}y_1 = \frac{2x_1y_1}{x_1 + y_1}.$$

As u_1 is true in the optimal assignment, the value of the optimal solution is $m = x_1$. Thus, for $k = 1$, the value of the expected performance ratio for the probabilistic greedy heuristic is

$$E(r_k) = \frac{2x_1y_1}{x_1(x_1 + y_1)} = \frac{2y_1}{x_1 + y_1} \leq 2.$$

Let $l \geq 1$ be an integer such that

$$E(r_k) \leq 2 \quad \text{for } k = l.$$

We show that

$$E(r_k) \leq 2 \quad \text{for } k = l + 1.$$

If the probabilistic greedy heuristic selects u_1 at Step 1, the value of the optimal solution at the second step of the greedy heuristic is $m - x_1$, where m is the optimal solution value of the k variable MINSAT problem P_k. However, if the greedy heuristic selects $\bar{u}_1$ at Step 1, the value of the optimal solution at the second step is bounded from above by m. Hence, the expected number of clauses satisfied by the probabilistic greedy heuristic is bounded from above by

$$q_1(x_1 + E(r_l)(m - x_1)) + (1 - q_1)(y_1 + E(r_l)m).$$

As $E(r_l) \leq 2$ by the induction hypothesis, the value of the above expression is no greater than

$$q_1(x_1 + 2(m - x_1)) + (1 - q_1)(y_1 + 2m).$$

Thus, an upper bound on the expected performance ratio for the probabilistic greedy heuristic is

$$E(r_{l+1}) \leq \frac{1}{m}(q_1(x_1 + 2(m - x_1)) + (1 - q_1)(y_1 + 2m))$$

$$= \frac{1}{m}(2m - q_1(x_1 + y_1) + y_1) = 2. \qquad \square$$

To prove the above bound is tight, consider the following example with $k = 2$ variables and $n = N + 1$ clauses. Let

$$c_1 = u_1 \vee u_2,$$

$$c_2 = \bar{u}_1,$$

$$c_i = \bar{u}_2, \qquad 3 \leq i \leq N + 1.$$

The optimal assignment sets both u_1 and u_2 true and satisfies one clause, c_1. The probabilistic greedy heuristic sets u_1 true or false with equal probability ($= \frac{1}{2}$) at its first step. If it sets u_1 true, then it obtains the optimal solution, setting u_2 true with probability 1 at its second step. Otherwise, at the second step, it

(i) sets u_2 true with probability $(N - 1)/N$, satisfying 2 clauses, c_1 and c_2, and

(ii) sets u_2 false with probability $1/N$, satisfying N clauses, $c_2, c_3, c_4, \ldots, c_{N+1}$.

Hence the expected performance ratio (= expected number of satisfied clauses) for the probabilistic greedy heuristic is

$$\frac{1}{2}\cdot 1 + \frac{1}{2}\left(\frac{N-1}{N}\cdot 2 + \frac{1}{N}\cdot N\right) = 2 - \frac{1}{N}.$$

As N tends to infinity, the value of this expression approaches from below the bound derived in Theorem 4. Note that $k = 2$ in this example and that the optimal clause c_1 contains $s = 2$ variables. Thus, unlike the worst-case and average performance bound for the (deterministic) greedy heuristic, the bound on the average performance of the probabilistic greedy heuristic does not depend on k or s.

5. Horn clauses. An important special case of the satisfiability problem occurs when each clause contains no more than one unnegated variable. Such clauses are called *Horn clauses*. The satisfiability problem defined over a set of Horn clauses can be solved in linear time (Dowling and Gallier [1]).

We show below that the MINSAT problem continues to be NP-hard even if it is restricted to a set of Horn clauses. In particular, we transform the 2-MINSAT problem to a MINSAT problem in Horn clauses.

Let the 2-MINSAT problem be defined over the set V of h variables and the set D of l clauses. We transform this to a MINSAT problem in Horn clauses defined over a set U of $k = h + l$ variables and a set C of $n = 3l$ clauses. Let $d = q_a \vee q_b$ be a clause in an instance of the 2-MINSAT problem, where q_a and q_b denote either variables in V or their negations. For each clause $d \in D$, define a variable w_d. Let $W = \{w_d | d \in D\}$. The set U is defined to be $V \cup W$. For each clause $d = q_a \vee q_b$ of 2-MINSAT, define three Horn clauses c_{1d}, $c_{2\mathrm{d}}$, and $c_{3d} \in C$, where

$$c_{1d} = q_a \vee \bar{w}_d, \quad c_{2d} = q_b \vee \bar{w}_d, \quad c_{3d} = w_d.$$

If clause d is satisfied by a truth assignment, then the same truth assignment for variables q_a and q_b, and a suitable truth assignment for variable w_d, satisfies two (the minimum that must be satisfied) of the above three clauses. If clause d is not satisfied by a truth assignment, the same truth assignment for variables q_a and q_b, and a suitable truth assignment for variable w_d satisfies one (the minimum that must be satisfied) of the above three clauses. Based on the above observations and using an argument similar to that in Theorem 1, we can show that no more than l^* clauses can be satisfied by some truth assignment for 2-MINSAT if and only if no more than $l + l^*$ clauses can be satisfied by a truth assignment for the MINSAT problem in Horn clauses.

We can verify that the worst-case and average-case bounds for the greedy heuristic and the probabilistic greedy heuristic remain unchanged for Horn clauses and that these bounds continue to be tight. Finally, we note that the MAXSAT problem in Horn clauses is also NP-hard (see the Appendix). Thus, while the satisfiability of Horn clauses can be assessed in linear time, the identification of assignments that maximize or minimize the number of Horn clauses satisfied are NP-hard problems.

Appendix. Complexity of MAXSAT for Horn clauses. We show below that the MAXSAT problem comprised of only Horn clauses is NP-hard. In particular, we transform the 2-MAXSAT problem to a MAXSAT problem in Horn clauses.

Let the 2-MAXSAT problem be defined over the set V of h variables and the set D of l clauses. We transform this problem to a MAXSAT problem in Horn clauses, defined over a set U of $k = h + 2l$ variables and a set C of $n = 5l$ clauses. Let $d = q_a \vee q_b$ be a clause in an instance of the 2-MAXSAT problem, where q_a and q_b denote either variables in V or their negations. For each clause $d \in D$, define two new variables w_{1d}, w_{2d}. Let $W = \{w_{1d}, w_{2d} | d \in D\}$. The set U is defined to be $V \cup W$. For each clause $d = q_a \vee q_b$ of 2-MAXSAT, define five Horn clauses $c_{1d}, c_{2d}, \ldots, c_{5d} \in C$, where

$$c_{1d} = q_a \vee \bar{w}_{1d},$$

$$c_{2d} = q_b \vee \bar{w}_{2d},$$

$$c_{3d} = \bar{w}_{1d} \vee \bar{w}_{2d},$$

$$c_{4d} = w_{1d},$$

$$c_{5d} = w_{2d}.$$

If clause d is satisfied by a truth assignment, the same truth assignment for variables q_a and q_b and a suitable truth assignment for variables w_{1d} and w_{2d} satisfy four (the maximum that can be satisfied) of the above five clauses in MAXSAT. If clause d is not satisfied by a truth assignment, the same truth assignment for variables q_a and q_b and a suitable truth assignment for variables w_{1d} and w_{2d} satisfy three (the maximum that can be satisfied) of the above five clauses. Based on the above observations and using an argument similar to that in Theorem 1, we can now show that no less than l^* clauses can be satisfied by some truth assignment for 2-MAXSAT if and only if no less than $3l + l^*$ clauses can be satisfied by a truth assignment for the MAXSAT problem in Horn clauses.

REFERENCES

[1] W. F. DOWLING AND J. H. GALLIER, *Linear-time algorithms for testing the satisfiability of propositional Horn formulae*, J. Logic Programming, 3 (1984), pp. 267–284.

[2] S. EVEN, A. ITAI, AND A. SHAMIR, *On the complexity of timetable and multicommodity flow problems*, SIAM J. Comput., 5 (1976), pp. 691–703.

[3] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractibility: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, CA, 1979.

[4] M. R. GAREY, D. S. JOHNSON, AND L. STOCKMEYER, *Some simplified* NP-*complete graph problems*, Theoret. Comput. Sci., 1 (1976), pp. 237–267.

[5] D. S. JOHNSON, *Approximation algorithms for combinatorial problems*, J. Comput. System Sci., 9 (1974), pp. 256–278.

[6] R. KOHLI AND R. KRISHNAMURTI, *Average performance of heuristics for satisfiability*, SIAM J. Discrete Math., 2 (1989), pp. 508–523.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 284-292, May 1994

014

ON OPTIMAL DEPTH THRESHOLD CIRCUITS FOR MULTIPLICATION AND RELATED PROBLEMS *

KAI-YEUNG SIU[†] AND VWANI P. ROYCHOWDHURY[‡]

Abstract. Let $\widehat{LT}_d$ denote the class of functions that can be computed by depth-d threshold circuits with polynomial size and polynomially bounded integer weights. Using the results in [M. Goldman, J. Håstad, and A. Razborov, in *Proc. 7th Annual Conference on Structure in Complexity Theory*], [M. Goldman and M. Karpinski, *Constructing depth $d+1$ majority circuits that simulate depth d threshold circuits*, unpublished] we show that multiple sum is in $\widehat{LT}_2$, and multiplication and division are in $\widehat{LT}_3$. Moreover, it follows from the lower-bound results in [A. Hajnal et al., *IEEE Sympos. Foundations of Comput. Sci.*, 28 (1987), pp. 99–110], [T. Hofmeister and P. Pudlák, *Forschungbericht Nr.* 477 Uni Dortmund, 1992] that these threshold circuits are optimal in circuit depth. The authors also indicate that these techniques can be applied to construct polynomial-size depth-3 threshold circuits for powering and depth-4 threshold circuits for multiple product.

Key words. threshold circuits, linear threshold functions, multiplication, division, arithmetic functions

AMS subject classifications. 68Q15, 68Q05, 68Rxx

1. Introduction. In this paper, we consider the power of small-depth threshold circuits in computing arithmetic functions such as multiple sum, multiplication, division, and powering. Threshold circuits are unbounded fan-in Boolean circuits in which each gate computes a *linear threshold function.* A linear threshold function $f(X)$ is a Boolean function such that

$$f(X) = \operatorname{sgn}(F(X)) = \begin{cases} 1 & \text{if } F(X) \geq 0, \\ 0 & \text{if } F(X) < 0, \end{cases}$$

where

$$F(X) = \sum_{i=1}^{n} w_i \cdot x_i + w_0.$$

The real coefficients w_i are commonly referred to as the *weights* of the threshold function. It is well known that the weights can be chosen to be integers [11].

However, the magnitudes of the integers can be exponentially large in the number of inputs. The *size* of a circuit is the number of gates. If the number of gates in a threshold circuit is polynomially bounded, then so is the number of wires in the circuit, and vice versa. Unless otherwise specified, we assume in the following discussions that all the weights in the threshold circuits are integers (possibly exponential) and that the sizes of the threshold circuits are polynomially bounded.

* Received by the editors April 1, 1992; accepted for publication (in revised form) November 22, 1992. This work was supported in part by the Joint Services Program at Stanford University (United States Army, Navy, and Air Force) contract DAAL03-88-C-0011 and the Department of the Navy (NAVELEX) contract N00039-84-C-0211, NASA Headquarters, Center for Aeronautics and Space Information Sciences grant NAGW-419-S6.

[†] Department of Electrical and Computer Engineering, University of California, Irvine, California 92717 (siu@balboa.eng.uci.edu). The work of this author was supported in part by Irvine Faculty Research grant 91/92-27.

[‡] School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907 (vwani@drum.ecn.purdue.edu). The work of this author was supported in part by the General Motors Faculty fellowship from the School of Engineering at Purdue University.

An important open problem in circuit complexity theory is to find an explicit function that cannot be computed in constant-depth polynomial-size threshold circuits. The first attempt toward solving this problem was made by Hajnal et al. [6]. They showed that the INNER PRODUCT MOD 2_n function $(x_1 \wedge y_1) \oplus \cdots \oplus (x_n \wedge y_n)$ can be computed in linear-size depth-3 threshold circuit but requires exponential-size depth-2 threshold circuit to compute, when the weights are polynomially bounded integers. This result gives a separation of the class of depth-2 polynomial-size threshold circuits from the class of depth-3 polynomial-size threshold circuits, when the weights are polynomially bounded.

Using the notation of [12], let us denote the class of depth-d polynomial-size threshold circuits where the weights are polynomially bounded by $\widehat{LT}_d$ and the corresponding class where the weights are unrestricted by LT_d. Then the exponential lower bound result on INNER PRODUCT MOD 2_n in [6] implies that $\widehat{LT}_2 \subsetneq \widehat{LT}_3$. As another consequence of this lower bound result, we can show that multiplication and division [9] of two n-bit integers also require exponential-size depth-2 threshold circuits with polynomially bounded weights to compute. In [7] Håstad and Goldmann proved an exponential lower bound on the size of depth-3 threshold circuits with the restriction that the bottom fan-in of the circuit is small. However, no exponential lower bound result on the size of depth-2 threshold circuits is known when there is no restriction on the size of the weights.

It is implicit in [3] that a constant-depth threshold circuit with arbitrary weights can be simulated by another constant-depth threshold circuit with polynomially bounded weights at the expense of at most a polynomial increase in size. To study the exact relationship between the depths of threshold circuits with arbitrary weights and with polynomially bounded weights, Siu and Bruck [12] showed that any depth-d threshold circuit can be simulated by a depth-$(2d+1)$ threshold circuit with polynomially bounded weights (both have polynomial size); that is, $LT_d \subset \widehat{LT}_{2d+1}$. The result was substantially strengthened by Goldmann, Håstad, and Razborov [4]; they showed that, in fact, $LT_d \subset \widehat{LT}_{d+1}$. While the proof techniques in [4] are not constructive, Goldmann and Karpinski [5] later gave an explicit construction of the circuits in [4].

Threshold circuits are powerful as a model of computation. In fact, many common arithmetic functions have been shown to be computable in small-depth threshold circuits. It was first shown in [12] that multiple sum and multiplication can be computed in $\widehat{LT}_3$ and $\widehat{LT}_4$, respectively. This result was also independently discovered later by Hofmeister, Hohberg, and Köhling [8] with much improvement on the circuit size. More recently, it was shown in [13] that small-depth threshold circuits can be constructed for division and related problems. In particular, division and powering are in $\widehat{LT}_4$ and multiple product (iterated multiplication) is in $\widehat{LT}_5$. The question of whether multiplication of two n-bit integers can be computed in $\widehat{LT}_3$ had remained open since the work of Hajnal et al. [6].

In this paper, we demonstrate that applications of the results in [4], [5] yield a depth-3 threshold circuit of polynomially bounded weights for multiplication; i.e., multiplication is in $\widehat{LT}_3$. It is clear from the result in [6] that such threshold circuit is optimal in depth. Moreover, similar techniques can be applied to show that division and powering can be computed in $\widehat{LT}_3$ and multiple product can be computed in $\widehat{LT}_4$. The result in [9] also implied that our division circuit is optimal in depth.

The rest of the paper is outlined as follows. We first describe a depth-2 threshold circuit for multiple sum. Using this result, a depth-3 threshold circuit for multiplication follows easily. We then indicate how to apply the result for multiple sum to obtain depth-3 threshold circuits for division and powering and depth-4 threshold circuit for multiple product. Since the techniques for deriving the results on division related problems are similar to those in [13], we only sketch the proof and indicate how the results in [13] can be improved using the results in [4], [5]. In the final section, we conclude with some open problems.

2. Main Results.

DEFINITION 1. *Given n n-bit integers,* $z_i = \sum_{j=0}^{n-1} z_{i,j}2^j$, $i = 1, \ldots, n$, $z_{i,j} \in \{0,1\}$, *we define* multiple sum *to be the problem of computing the* $(n + \log n)$*-bit sum* $\sum_{i=1}^{n} z_i$ *of the n integers.*

The above problem is also referred to as *iterated addition* in the literature.

DEFINITION 2. *Given two n-bit integers,* $x = \sum_{j=0}^{n-1} x_j 2^j$ *and* $y = \sum_{j=0}^{n-1} y_j 2^j$, *we define* multiplication *to be the problem of computing the* $(2n)$*-bit product of x and y.*

It is easy to see that, if multiple sum can be computed in $\widehat{LT}_2$, then multiplication can be computed in $\widehat{LT}_3$. We first prove the result on multiple sum. Our result hinges on the results in [4], [5]. The key observation is that multiple sum can be computed as a sum of polynomially many linear threshold (LT_1) functions (with exponential weights). Let us first state the results [4], [5].

LEMMA 2.1 (see [4], [5]). *Let* $\widetilde{LT}_d$ *denote the class of depth-d polynomial-size threshold circuits where the weights at the output gate are polynomially bounded integers (with no restriction on the weights of the other gates). Then* $\widetilde{LT}_d = \widehat{LT}_d$ *for any fixed integer* $d \geq 1$.

The following lemma is a generalization of a result in [10]. Informally, the result says that, if a function is 1 when a weighted sum (possibly exponential) of its inputs lies in one of polynomially many intervals, and is 0 otherwise, then the function can be computed as a sum of polynomially many LT_1 functions.

LEMMA 2.2. *Let* $S = \sum_{i=1}^{n} w_i x_i$ *and* $f(X)$ *be a function such that* $f = 1$ *if* $S \in [l_i, u_i]$ *for* $i = 1, \ldots, N$ *and* $f = 0$ *otherwise, where N is polynomially bounded in n. Then f can be computed as a sum of polynomially many* LT_1 *functions, and thus* $f \in \widetilde{LT}_2$.

Proof. For $j = 1, \ldots, N$, let

$$y_{l_j} = \operatorname{sgn}\left\{\sum_{i=1}^{n} w_i x_i - l_j\right\} \quad \text{and} \quad y_{u_j} = \operatorname{sgn}\left\{u_j - \sum_{i=1}^{n} w_i x_i\right\}.$$

We claim that

$$f(X) = \sum_{j=1}^{N}(y_{l_j} + y_{u_j}) - N,$$

and therefore

$$f(X) = \operatorname{sgn}\left\{\sum_{j=1}^{N}(y_{l_j} + y_{u_j}) - N - 1\right\}.$$

Note the following: If, for $j = 1, \ldots, N$, $\sum_{i=1}^{n} w_i x_i \notin [l_j, u_j]$, then $y_{l_j} + y_{u_j} = 1$ for all j. Thus, $\sum_{j=1}^{N}(y_{l_j} + y_{u_j}) - N = 0$. On the other hand, if $\sum_{i=1}^{n} w_i x_i \in [l_j, u_j]$ for some $j \in \{1, \ldots, N\}$, then $y_{l_j} + y_{u_j} = 2$ and $y_{l_i} + y_{u_i} = 1$ for $i \neq j$. Thus, $\sum_{i=1}^{N}(y_{l_i} + y_{u_i}) - N = N + 1 - N = 1$. □

Combining the above two lemmas yields a depth-2 threshold circuit for multiple sum.

THEOREM 2.3. *Multiple sum is in* $\widehat{LT}_2$.

Proof. Given n n-bit integers $z_i = \sum_{j=0}^{n-1} z_{i,j}2^j$, $i = 1, \dots, n$, the sum $\tilde{S} = \sum_{i=1}^{n} z_i$ can be represented as an $(n + \log n)$-bit integer, $\tilde{S} = \sum_{i=0}^{n+\log n-1} \tilde{s}_i 2^i$. Clearly, the kth bit of $\tilde{S}$, $\tilde{s}_{k-1}$ is the same as the kth bit of the sum of the first k-bits of the z_i's, i.e., the kth bit of $\sum_{i=1}^{n}\sum_{j=0}^{k-1} z_{i,j}2^j$. Thus, to prove the theorem, it suffices to show that the kth bit of n k-bit integers can be computed in $\widehat{LT}_2$, for $k = 1, \dots, n + \log n$. We first construct a depth-2 threshold circuit where the threshold gates in the first level have exponential weights.

Let $S = \sum_{i=1}^{n}\sum_{j=0}^{k-1} 2^j z_{i,j} = \sum_{l=0}^{\log n+k-1} 2^l s_l$ be the sum of n k-bit integers. Note that the kth bit of S, s_{k-1} is 1 if $S \in I_{j,k} = [j2^{k-1}, (j+1)2^{k-1} - 1]$ for $j = 1, 3, 5, \dots, 2^{\log n+1} - 1$ and 0 otherwise. Since there are only polynomially many intervals $I_{j,k}$, it follows from Lemma 2.2 the kth bit can be computed in $\widetilde{LT}_2$. Now apply Lemma 2.1 for $d = 2$; thus the kth bit can be computed in $\widehat{LT}_2$. □

It is also easy to see that multiple sum cannot be computed in LT_1. Simply observe that the first bit of the sum is the parity function, which does not belong to LT_1. Thus the above threshold circuit for multiple sum has minimum possible depth.

THEOREM 2.4. *Multiplication is in* $\widehat{LT}_3$.

Proof. Let the two integers be $x = x_{n-1}x_{n-2}\cdots x_0$, $y = y_{n-1}y_{n-2}\cdots y_0$. The first level of our circuit outputs the n $(2n)$-bit integers $z_i = z_{i_{2n-1}}z_{i_{2n-2}}\cdots z_{i_0}$, for $i = 0, \dots, n-1$, where

$$z_i = \underbrace{0\dots0}_{n-i}(x_{n-1} \wedge y_i)(x_{n-2} \wedge y_i)\dots(x_0 \wedge y_i)\underbrace{0\dots0}_{i}.$$

This level requires $O(n^2)$ gates. It is easy to see that the product of x and y is simply the sum of the z_i's. By Theorem 2.3, the sum of the z_i's can be computed using two more levels of polynomially many threshold gates (with polynomially bounded weights). □

We can further apply the results in [4], [5] to construct small-depth threshold circuits for division, powering, and multiple product. Let us give a formal definition of these problems.

DEFINITION 3. *Let Z be an n-bit integer ≥ 0. We define* powering *to be the n^2-bit representation of Z^n.*

DEFINITION 4. *Given n n-bit integers z_i, $i = 1, \dots, n$, we define* multiple product *to be the n^2-bit representation of $\prod_{i=1}^{n} z_i$.*

The above problem is also called *iterated product* or *iterated multiplication* in the literature.

Suppose that we want to compute the quotient of two integers. Some quotient in binary representation might require infinitely many bits; however, a circuit can only compute the most significant bits of the quotient. If a number has both finite and infinite binary representation (for example, 0.1 = 0.0111...), we always express the number in its finite binary representation. We are interested in computing the truncated quotient, defined below.

DEFINITION 5. *Let X and $Y \geq 1$ be two input n bit integers. Let $X/Y = \sum_{i=-\infty}^{n-1} z_i 2^i$ be the quotient of X divided by Y. We define* $\mathrm{DIV}_k(X/Y)$ *to be X/Y*

truncated *to the* $(n+k)$*-bit number, i.e.,*

$$\mathrm{DIV}_k(X/Y) = \sum_{i=-k}^{n-1} z_i 2^i.$$

In particular, $\mathrm{DIV}_0(X/Y)$ is $\lfloor X/Y \rfloor$, the greatest integer $\leq X/Y$.

Using the results in [2], [12], it was shown in [13] that powering and division can be computed in $\widehat{LT}_4$ and that multiple product can be computed in $\widehat{LT}_5$. Combining these results with the results in [4], [5], we can reduce the depths of these circuits by one. We only indicate the key steps in the construction of the circuits in [13]. For other details of the proof, see [13]. Let us rephrase the results in [4], [5] as the following lemma.

LEMMA 2.5 (see [4], [5]). *Let* $f(X) \in LT_1$. *Then, for any* $k > 0$, *there exist* m *functions* $t_1(X), \ldots, t_m(X) \in \widehat{LT}_1$ *such that, for all* X,

$$\left| f(X) - \frac{1}{N} \sum_{j=1}^{m} t_j(X) \right| \leq n^{-k},$$

where m *and* N *are integers bounded by a polynomial in* n.

By Lemma 2.2, each bit in the sum of multiple sum can be computed as a sum of polynomially many LT_1 functions. Combining this result with the above lemma yields the following result.

LEMMA 2.6. *Let* s_i *be any of the outputs in multiple sum. Then, for any* $\hat{k} > 0$, *there exist* $\hat{t}_j(X) \in \widehat{LT}_1$ *such that*

$$\left| s_i - \frac{1}{\hat{N}} \sum_{j=1}^{\hat{m}} \hat{t}_j(X) \right| \leq n^{-\hat{k}},$$

where $\hat{m}$ *and* $\hat{N}$ *are integers bounded by a polynomial in* n.

The following lemma, which was shown in [13], states that, if t_1 and t_2 can be closely approximated by polynomially many $\widehat{LT}_k$ functions, so is their product $t_1 \wedge t_2$.

LEMMA 2.7. *Suppose that, for* $i = 1, 2$ *and for every* $c > 0$, *there exist integers* m_i, w_{i_j}, *and* N *that are bounded by a polynomial in* n *such that, for all inputs* X,

$$\left| t_i(X) - \frac{1}{N} \sum_{j=1}^{m_i} w_{i_j} t_{i_j}(X) \right| = O(n^{-c}),$$

where each $t_{i_j} \in \widehat{LT}_k$. *Then there exist integers* $\tilde{m}$, $\tilde{w}_j$, *and* $\tilde{N}$ *that are bounded by a polynomial in* n *such that*

$$\left| t_1(X) \wedge t_2(X) - \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{m}} \tilde{w}_j \tilde{t}_j(X) \right| = O(n^{-c}),$$

where each $\tilde{t}_j \in \widehat{LT}_k$.

To avoid cumbersome explanations in the following discussions, we say informally that every LT_1 function and each output bit in multiple sum can be *closely approximated* by a sum of polynomially many $\widehat{LT}_1$ functions, in the sense of Lemmas 2.5 and 2.6.

THEOREM 2.8. *Powering is in* $\widehat{LT}_3$.

Proof. Let X be an n-bit integer and $Z = X^n$. Let p_i denote the ith prime number and let $\pi(k)$ denote the number of primes $\leq k$. Let

$$P_n = \prod_{i=1}^{\pi(n^2)} p_i$$

be the product of all primes $\leq n^2$. Then we can show that $Z < 2^{n^2} < P_n$, and thus $(Z \bmod P_n) = Z$.

Using the Chinese Remainder Theorem, we can compute Z with the following steps:

1. For $i = 1, ..., n^2$, compute in parallel the values $r_i = Z \bmod p_i$;
2. $\tilde{Z} = \sum_{i=1}^{\pi(n^2)} r_i \cdot m_i$;
3. $Z = (Z \bmod P_n) = (\tilde{Z} \bmod P_n)$;

where, in step 2 above, the m_i are fixed integers (possibly exponentially large), and therefore step 2 is, in fact, multiple sum. Moreover, we can show that $\tilde{Z} \leq n^4 \cdot P_n$, and hence $Z = (\tilde{Z} \bmod P_n) = \tilde{Z} - k \cdot P_n$ for some k, where $0 \leq k \leq n^4$. For each $k \in \{0, \ldots, n^4\}$, let

$$\begin{aligned} EQ_k(Z) &= \operatorname{sgn}\{\tilde{Z} - k \cdot P_n\} + \operatorname{sgn}\{(k+1)P_n - \tilde{Z} - 1\} - 1 \\ &= \begin{cases} 1 & \text{if } Z = (\tilde{Z} \bmod P_n) = \tilde{Z} - k \cdot P_n, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Let z_{jk} be the jth bit of $\tilde{Z} - k \cdot P_n$. Then the jth bit of Z is

$$\bigvee_{0 \leq k \leq n^4} (EQ_k(Z) \wedge z_{jk}).$$

We can compute the values r_i in step 1, above, as a sum of polynomially many $\widehat{LT}_1$ functions. By Lemmas 2.6 and 2.5, each z_{jk} and each $EQ_k(Z)$ can be *closely approximated* by a sum of polynomially many $\widehat{LT}_1$ functions with variables r_i. Thus $EQ_k(Z)$ and z_{jk} can be *closely approximated* as a sum of the outputs from polynomially many depth-2 threshold circuits whose inputs are the variables X. By Lemma 2.7, it follows that $(EQ_k(Z) \wedge z_{jk})$ can also be closely approximated as a sum of the outputs from polynomially many depth-2 threshold circuits. Hence, each of the outputs $\bigvee_{0 \leq k \leq n^4}(EQ_k(Z) \wedge z_{jk})$ can be computed in a depth-3 threshold circuit. □

Remark 1. In [13] each $EQ_k(Z)$ and z_{jk} is closely approximated as a sum of outputs from polynomially many depth-3 threshold circuits ($\widehat{LT}_3$). Lemmas 2.6 and 2.5 enable us to save one level of threshold gates in computing them.

THEOREM 2.9. *Multiple product is in* $\widehat{LT}_4$.

Proof. Let $Z = \prod_{j=1}^{n} z_j$, where each z_i is an n-bit integer. The proof is very similar to the proof of Theorem 2.8. We can compute Z using the same three steps as in Theorem 2.8. The only difference is that now each $r_i = Z \bmod p_i$ is computed as a sum of polynomially many depth-2 threshold circuits ($\widehat{LT}_2$), one more level of threshold gates than the circuit for powering. □

THEOREM 2.10. $\mathrm{DIV}_k(x/y)$ *is in* $\widehat{LT}_3$.

Proof. Note that $\mathrm{DIV}_k(x/y) = 2^{-k}\mathrm{DIV}_0(2^k x/y)$; so it suffices to prove our claim for the case where $k = 0$. The resulting threshold circuits for the general case when

k is polynomial in n have the same depth and the size will increase by a polynomial factor.

The underlying idea is to compute an *overapproximation* $\tilde{a}$ to x/y such that $x/y \leq \tilde{a} \leq x/y + 2^{-(n+1)}$. Then we can show that $\lfloor \tilde{a} \rfloor = \lfloor x/y \rfloor$.

Since x/y is equal to the product of x and y^{-1}, it is enough to get an overapproximation $\tilde{y}^{-1}$ of y^{-1} with error $\leq 2^{-(2n+1)}$. Then we can compute the approximation $q = x \cdot \tilde{y}^{-1}$ to x/y with an error $\leq x2^{-(2n+1)} \leq 2^{-(n+1)}$ with a small-depth threshold circuit.

To construct an overapproximation of y^{-1}, let $j \geq 1$ be the integer such that $2^{j-1} \leq y < 2^j$. Note that $|1 - y2^{-j}| \leq \frac{1}{2}$, and we can express y^{-1} as a series expansion

$$y^{-1} = 2^{-j} \cdot (1 - (1 - y2^{-j}))^{-1} = 2^{-j} \sum_{i=0}^{\infty} (1 - y2^{-j})^i.$$

If we put

$$\tilde{y}^{-1} = 2^{-j} \sum_{i=0}^{2n} (1 - y2^{-j})^i,$$

then the difference

$$0 \leq (y^{-1} - \tilde{y}^{-1}) \leq 2^{-j} \sum_{i=(2n+1)}^{\infty} 2^{-i} \leq 2^{-(2n+1)}.$$

Since $x < 2^n$, we have

$$0 \leq (xy^{-1} - x\tilde{y}^{-1}) < 2^{-(n+1)}.$$

Suppose for the moment that we can find the integer $j \geq 1$ such that $2^{j-1} \leq y < 2^j$. Now we can rewrite

$$x\tilde{y}^{-1} = \frac{1}{2^{j(2n+2)}} \sum_{i=0}^{2n+1} 2^{j(2n+1-i)} x(2^j - y)^i.$$

Let $Z_j = \sum_{i=0}^{2n+1} 2^{j(2n+1-i)} x(2^j - y)^i$. Then $x\tilde{y}^{-1} = (1/2^{j(2n+2)})Z_j$, a shifting of the bits in Z_j.

Again, we can compute Z_j via the Chinese Remainder Theorem as follows:

1. For $i = 1, \ldots, N$, compute in parallel the values $r_{i,j} = Z_j \bmod p_i$;
2. $\tilde{Z}_j = \sum_{i=1}^{N} r_{i,j} \cdot m_i$;
3. $Z_j = (Z_j \bmod P_N) = (\tilde{Z}_j \bmod P_N)$;

where N is a sufficiently large integer such that the product of the first N primes $\prod_{i=1}^{N} p_i = P_N > Z_j$ for all $j = 1, \ldots, n$. Moreover, we can show that $\tilde{Z}_j \leq n^\alpha P_N$ for some $\alpha > 0$, and hence $Z_j = (\tilde{Z}_j \bmod P_N) = \tilde{Z}_j - kP_n$ for some k, where $0 \leq k \leq n^\alpha$. For each $k \in \{0, ..., n^\alpha\}$, let

$$\begin{aligned} EQ_k(Z_j) &= \operatorname{sgn}\{\tilde{Z}_j - kP_N\} + \operatorname{sgn}\{(k+1)P_N - \tilde{Z}_j - 1\} - 1 \\ &= \begin{cases} 1 & \text{if } Z_j = (\tilde{Z}_j \bmod P_N) = \tilde{Z}_j - kP_N, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Let $\sum_l z_{j,k,l} 2^l = (1/2^{j(2n+2)})(\tilde{Z}_j - kP_N)$. If $EQ_{k^*}(Z_j) = 1$, i.e., $(\tilde{Z}_j \bmod P_N) = \tilde{Z}_j - k^* P_N$, then

$$\mathrm{DIV}_0(x/y) = \sum_{l=0}^{n-1} z_{j,k^*,l} 2^l.$$

Thus the ith bit of $\mathrm{DIV}_0(x/y)$ can be computed as

$$\bigvee_{1 \le k \le n^\alpha} (EQ_k(Z_j) \wedge z_{j,k,i}).$$

The above expression is based on the assumption that we can find the unique integer $j \geq 1$ such that $2^{j-1} \leq y < 2^j$. We can compute such integer j in parallel without increasing the depth of the circuit. To see this, for each $j \in \{1, \ldots, n\}$, let

$$I_j = \mathrm{sgn}\{y - 2^{j-1}\} + \mathrm{sgn}\{2^j - y - 1\} - 1 = \begin{cases} 1 & \text{if } 2^{j-1} \leq y < 2^j, \\ 0 & \text{otherwise.} \end{cases}$$

Then the ith bit of $\mathrm{DIV}_0(x/y)$ is

$$\bigvee_{1 \le j \le n} \bigvee_{1 \le k \le n^\alpha} (I_j \wedge EQ_k(Z_j) \wedge z_{j,k,i}).$$

Now apply the same argument as in Theorem 2.8; we can show that each $(I_j \wedge EQ_k(Z_j) \wedge z_{j,k,i})$ can be closely approximated by a sum of outputs from polynomially many $\widehat{LT}_2$ functions. Hence the final result can be computed in $\widehat{LT}_3$. □

Remark 2. In [13] each $(I_j \wedge EQ_k(Z_j) \wedge z_{j,k,i})$ is closely approximated by a sum of outputs from polynomially many depth-3 threshold circuits $(\widehat{LT}_3)$. Here again Lemmas 2.6 and 2.5 enable us to save one level of threshold gates in computing them.

3. Concluding remarks. We have demonstrated optimal-depth threshold circuits for multiplication, multiple sum, and division. We also indicated how the techniques can be applied to obtain depth-3 threshold circuits for powering and depth-4 threshold circuit for multiple product. These results are improvements on the depths of the circuits constructed in [13]. Moreover, the construction of these circuits can all be made explicit using the results in [5].

There are a few open problems, below, related to the results in this paper:

1. What is the minimal size of a depth-3 threshold circuit for multiplication?
2. Can INNER PRODUCT MOD 2_n and multiplication be computed in LT_2? A negative answer to this question will provide the separation $LT_2 \subsetneq \widehat{LT}_3$.

REFERENCES

[1] N. Alon and J. Bruck, *Explicit Constructions of Depth-2 Majority Circuits for Comparison and Addition*, IBM Research Report, RJ 8300, August 1991.

[2] P. W. Beame, S. A. Cook, and H. J. Hoover, *Log depth circuits for division and related problems*, SIAM J. Comput., 15 (1986), pp. 994–1003.

[3] A. K. Chandra, L. Stockmeyer, and U. Vishkin, *Constant depth reducibility*, SIAM J. Comput., 13 (1984), pp. 423–439.

[4] M. Goldmann, J. Håstad, and A. Razborov, *Majority gates vs. general weighted threshold gates*, in Proc. 7th Annual Conference on Structure in Complexity Theory Conference, 1992, pp. 2–13.

[5] M. Goldmann and M. Karpinski, *Simulating threshold circuits by majority circuits*, in Proc. of the 25th annual ACM Symposium on the Theory of Computing (STOC), San Diego, CA, May 1993, pp. 551–560.

[6] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, and G. Turán, *Threshold circuits of bounded depth*, IEEE Sympos. Foundations of Comput. Sci., 28 (1987), pp. 99–110.

[7] J. Håstad and M. Goldmann, *On the power of small-depth threshold circuits*, Computational Complexity, 1 (1991), pp. 113-129.

[8] T. Hofmeister, W. Hohberg, and S. Köhling, *Some notes on threshold circuits and multiplication in depth* 4, Inform. Process. Lett., 39 (1991), pp. 219–225.

[9] T. Hofmeister and P. Pudlák, *A proof that division is not in* TC_2^0, Forschungsbericht Nr. 447, Uni Dortmund, 1992.

[10] S. Muroga, *The principle of majority decision logic elements and the complexity of their circuits*, Internat. Conf. on Information Processing, Paris, France, June 1959.

[11] S. Muroga, I. Toda, and S. Takasu, *Theory of majority decision elements*, J. Franklin Inst., 271 (1961), pp. 376–418.

[12] K.-Y. Siu and J. Bruck, *On the power of threshold circuits with small weights*, SIAM J. Discrete Math., 4 (1991), pp. 423–435.

[13] K.-Y. Siu, J. Bruck, T. Kailath, and T. Hofmeister, *Depth-efficient neural networks for division and related problems*, IEEE Trans. Inform. Theory, 39 (1933), pp. 946–956.

ON FINDING CRITICAL INDEPENDENT AND VERTEX SETS*

ALEXANDER A. AGEEV[†]

Abstract. An independent set I_c of a undirected graph G is called critical if

$$|I_c| - |N(I_c)| = \max\{|I| - |N(I)| : I \text{ is an independent set of } G\},$$

where $N(I)$ is the set of all vertices of G adjacent to some vertex of I. It has been proved by Cun-Quan Zhang [*SIAM J. Discrete Math.*, 3 (1990), pp. 431–438] that the problem of finding a critical independent set is polynomially solvable. This paper shows that the problem can be solved in $O(|V(G)|^{1/2}|E(G)|)$ time and its weighted version in $O(|V(G)|^2|E(G)|^{1/2})$ time.

Key words. independent set, minimum cut

AMS subject classifications. 05C35, 68R10

1. Introduction. Denote by G a simple undirected graph and by $N(U)$, $U \subseteq V(G)$, the set of all vertices of G adjacent to some vertex of U. An independent set $I_c \subseteq V(G)$ is called *critical* if

$$\nu(G) = |I_c| - |N(I_c)| = \max\{|I| - |N(I)| : I \text{ is an independent set of } G\}.$$

The number $\nu(G)$ is a parameter of G, closely related to some other important ones [Zh90]. A vertex set U_c is called *critical* if

$$\mu(G) = |U_c| - |N(U_c)| = \max\{|U| - |N(U)| : U \subseteq V(G)\}.$$

Cun-Quan Zhang [Zh90] observed that $\nu(G) = \mu(G)$ and that the problem of finding critical independent set is reducible to the problem of finding critical vertex set. In [Zh90] it is also shown by a rather sophisticated reduction to the linear program that the problem of finding critical vertex set is polynomially solvable. This paper presents a very simple reduction of the weighted version of the problem to the maximum weight independent set problem on bipartite graphs. As a consequence, we obtain a fast algorithm to find critical independent and vertex sets in general graphs.

2. Critical weighted sets. Let G be a undirected graph with nonnegative weights $w(v)$ on its vertices. For any $X \subseteq V$, denote $\rho(X) = w(X) - w(N(X))$, where $w(S)$ is the total weight $\sum_{v \in S} w(v)$ of $S \subseteq V$.

A vertex set $U^* \subseteq V(G)$ is *critical weighted* if $\rho(U^*) = \max\{\rho(U) : U \subseteq V(G)\}$. An independent set I^* is *critical weighted* if

$$\rho(I^*) = \max\{\rho(I) : I \text{ is an independent set of } G\}.$$

Obviously, $\rho(I^*) \le \rho(U^*)$. We claim that $\rho(I^*) = \rho(U^*)$. Indeed, let U be a critical weighted vertex set. Define $I \subseteq U$ to be the set of all isolated vertices of the subgraph induced by U. Then, clearly, I is an independent set of G. Since $N(U) = U_1 \cup X$, where $U_1 = U \setminus I$, $X = (N(U_1) \setminus U_1) \cup N(I)$, we have

$$\rho(U) = w(U_1) + w(I) - w(U_1) - w(X) \le \rho(I).$$

It follows that $\rho(U) = \rho(I)$ and I is a critical weighted independent set, of G.

Thus, to find a critical weighted independent set, it suffices to find a critical weighted vertex set.

* Received by the editors June 10, 1991; accepted for publication (in revised form) December 10, 1992.

[†] Institute of Mathematics, Universitetskii pr. 4, Novosibirsk 630090, Russia (ageev@math.nsk.su).

3. The reduction. Suppose that we have an instance of the problem of finding critical weighted vertex set. Let (x_v), (y_v) be $(0,1)$ vectors whose components are indexed by the vertices of G. Consider the following $(0,1)$ programming problem:

$$\max \sum_{u\in V} w(u)x_u - \sum_{v\in V} w(v)y_v, \tag{1}$$

$$\text{s.t. } y_v \geq x_u, \qquad (u,v) \in E(G), \tag{2}$$

$$x_u, y_v \in \{0,1\}, \qquad u,v \in V(G). \tag{3}$$

We claim that, if (x_v^*), (y_v^*) is an optimal solution to this problem, then (x_v^*) is the incidence vector of a critical weighted vertex set. Indeed, let $X^* = \{v \in V : x_v^* = 1\}$. Observe that (y_v^*) is the incidence vector of $N(X^*)$. Now let (x_v) be the incidence vector of some $X \subseteq V$ and let (y_v) be the incidence vector of $N(X)$. Since $(x_v), (y_v)$ is a feasible solution of (1)–(3), it follows that

$$\rho(X) = \sum_{u\in V} w(u)x_u - \sum_{v\in V} w(v)y_v \leq \sum_{u\in V} w(u)x_u^* - \sum_{v\in V} w(v)y_v^* = \rho(X^*).$$

It is easy to see that (1)–(3) is an instance of the selection problem [Ba70], [Rh70]. Finding an optimal solution of (1)–(3) is known to be equivalent to finding a minimum cut in a bipartite network [Ba70], [Rh70], [PQ82]. Hence an optimal solution of (1)–(3) can be found in $O(|V(G)|^2|E(G)|^{1/2})$ time [GT88], [CM89]. Furthermore, putting $y_v = 1 - z_v$, $v \in V$, we can rewrite (1)–(3) as follows:

$$\max \sum_{u\in V} w(u)x_u + \sum_{v\in V} w(v)z_v + |V(G)|,$$

$$\text{s.t. } x_u + z_v \leq 1, \qquad (u,v) \in E(G),$$

$$x_u, z_v \in \{0,1\}, \qquad u,v \in V(G).$$

The latter is nothing but an instance of the maximum weight independent set problem on bipartite graph H with bipartition (V, V'), where V' is a copy of V and for any $u \in V$, $v' \in V'$,

$$(u,v') \in E(H) \quad \text{if and only if} \quad (u,v) \in E(G).$$

In the case of unit weights $(w(v) \equiv 1)$, an optimal solution to this can be obtained as a by-product with the $O(|V(G)|^{1/2}|E(G)|)$ algorithm for finding maximum matching in bipartite graphs [PS82, p. 226].

REFERENCES

[Ba70] M. L. Balinski, *On a selection problem*, Management Sci., 17 (1970), pp. 230–231.

[CM89] J. Cheriyan and S. N. Maheshwary, *Analysis of preflow push algorithms for maximum network flow*, SIAM J. Comput., 18 (1989), pp. 1057–1086.

[GT88] A. V. Goldberg and R. E. Tarjan, *A new approach to the maximum flow problem*, J. Assoc. Comput. Mach., 35 (1988), pp. 921–940.

[PS82] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[PQ82] J.-C. Picard and M. Queyranne, *Selected applications of minimum cuts in networks*, INFOR, 20 (1982), pp. 394–422.

[Rh70] J. M. W. Rhys, *A selection problem of shared fixed costs and network flows*, Management Sci., 17 (1970), pp. 200–207.

[Zh90] C.-Q. Zhang, *Finding critical independent sets and critical vertex subsets are polynomial problems*, SIAM J. Discrete Math., 3 (1990), pp. 431–438.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 296–306, May 1994

016

TRIANGULATING VERTEX-COLORED GRAPHS*

F. R. MCMORRIS[†], TANDY J. WARNOW[‡], AND THOMAS WIMER[§]

Abstract. This paper examines the class of vertex-colored graphs that can be triangulated without the introduction of edges between vertices of the same color. This is related to a fundamental and long-standing problem for numerical taxonomists, called the Perfect Phylogeny Problem. These problems are known to be polynomially equivalent and NP-complete. This paper presents a dynamic programming algorithm that can be used to determine whether a given vertex-colored graph can be so triangulated and that runs in $O((n + m(k-2))^{k+1})$ time, where the graph has n vertices, m edges, and k colors. The corresponding algorithm for the Perfect Phylogeny Problem runs in $O(r^{k+1}k^{k+1} + sk^2)$ time, where s species are defined by k r-state characters.

Key words. graph algorithms, evolution, k-trees

AMS subject classifications. 05C85, 68Q20, 68Q25, 92B10, 05C05

1. Introduction.

We are interested in two problems. The first, which has a purely graph-theoretic statement, is as follows: Given a graph $G = (V, E)$ with a proper vertex coloring, $c : V \rightarrow \{1, 2, \ldots, k\}$, we wish to determine whether we can add edges to G so as to make it chordal, without introducing edges between vertices of the same color. A *chordal* or *triangulated* graph is a graph that has no induced cycles of length 4 or more. If we can triangulate G in this way, we say that G has a c-triangulation, or that G is *c-triangulatable.* We call this the *Triangulating Colored Graphs Problem*, or TCGP.

The second problem is a fundamental and long-standing problem for numerical taxonomists. This problem, called the *Perfect Phylogeny Problem*, or PPP, is as follows. A *phylogeny* is a rooted tree that describes the evolution of a set S of species. The species in S are at the leaves of the tree, and the internal nodes represent ancestral species. In this problem, each species is defined in terms of characters, both the species at the leaves (which represent extant species) and the species at the internal nodes (which represent ancestral species). For example, a character can be binary, such as *vertebrate-invertebrate*, or it can take several states, such as the character that has a state for each possible number of legs in a species. Thus, the phylogeny is a rooted tree, in which each vertex is labelled with a k-vector of character states. When the phylogeny has the property that each state is transited into at most once, so that, for each character state, the nodes having that character state form a subtree, we say that the character set is compatible, and the phylogeny is said to be perfect. The PPP is therefore as follows: Given a set S of n species and a set C of k characters defined on the set S, determine whether a perfect phylogeny exists for the species and characters. This is also known as the *Character Compatibility Problem.* A thorough discussion of the biological setting of this problem can be found in [8].

In 1974, Buneman [5] proved that the PPP reduces in polynomial time to the TCGP. The reduction from the PPP to TCGP describes each input I to the PPP as

* Received by the editors April 13, 1992; accepted for publication (in revised form) December 16, 1992.

[†] Department of Mathematics, University of Louisville, Louisville, Kentucky 40292. The work of this author was supported by Office of Naval Research grant N00014-89-J-1643.

[‡] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104. The work of this author was supported by National Science Foundation grants CCR88-13632 and DMS9005833, and by U.S. Department of Energy contract DE-AC04-76DP00789.

[§] Department of Mathematics, Clarion University, Clarion, Pennsylvania 16214.

a vertex-colored graph G_I. The vertices in G_I are labelled by the distinct character states in I, with vertices corresponding to states of the same character given the same color. If two states share at least one species in common, the corresponding vertices in G_I are made adjacent. Thus, the graph G_I has the property that it is properly colored by k colors and is edge-covered by k-cliques, where k is the number of characters in I. Graphs having this property are called *partition intersection graphs* [14]. Buneman's theorem then says that a perfect phylogeny exists for I if and only if the graph G_I can be c-triangulated.

Very recently, results have shown TCGP and PPP to be polynomially equivalent [19] and NP-complete [3], [18]. For a long time, the only solutions were for binary (two-state) characters [6], [10], [20] and for two characters at a time [7], [9]. An algorithm for the PPP has been found when the characters are restricted to having at most four states [13]; this algorithm can be used to construct perfect phylogenies from DNA sequences. The TCGP has been solved for three-colored graphs in [12], [11], [4].

For $k > 3$, the only result relevant to triangulating k-colored graphs is the negative result of Bodlaender, Fellows, and Warnow, who showed that for $k > 3$, the class of k-colored graphs that can be c-triangulated is not *finite-state* [3]. The significance of this last result is the following: There is a growing body of literature that shows that many NP-hard problems become solvable in linear time when applied to graphs of bounded treewidth, given a suitable tree-decomposition, when the class of graphs is *finite-state.* One of the results of this paper is that a c-triangulatable k-colored graph has treewidth bounded by $k - 1$; if this class of c-triangulatable k-colored graphs was finite-state, we would be able to obtain a linear time algorithm to recognize these graphs.

To say that a graph $G = (V, E)$ has treewidth bounded by k (i.e., is a partial k-tree) is to assert that a supergraph $G' = (V, E')$ of G exists that is a k-tree. We were thus motivated to examine the known results about partial k-tree recognition and, in particular, the $O(n^{k+2})$ time dynamic programming algorithm of Arnborg, Corneil, and Proskurowski [1] for recognizing partial k-trees. However, the recognition of c-triangulatable $(k-1)$-colored graphs is more difficult than recognizing partial k-trees, for the following reason: Each k-tree containing at least $k + 1$ vertices contains a $(k + 1)$-clique; thus, while we may be able to embed the $(k + 1)$-colored graph in a properly colored k-tree, we will not be able to embed it in a properly colored k'-tree, for any $k' > k$. Contrary to this is the fact that, if a graph G is a partial k-tree, it is also a partial k'-tree, for every $k' > k$. Thus, the ability to use a dynamic programming approach for partial k-tree recognition is not immediately transferrable to recognizing c-triangulatable $(k - 1)$-colored graphs. Nevertheless, we were able to obtain an $O((n + m(k - 2))^{k+1})$ algorithm to determine whether a k-colored graph with n vertices and m edges can be c-triangulated. This algorithm can be used to determine the compatibility of k r-state characters on s species in $O(r^{k+1}k^{k+1} + sk^2)$ time. Thus, when we fix the number of colors (characters), TCGP (PPP) is in P.

2. Definitions. A *graph* $G = (V, E)$ is a finite undirected graph without multiple edges or loops. We say that a graph $G = (V, E)$ is *connected* if every two distinct vertices in V are connected by a path. The maximal subgraphs of G that are connected are called the *components* of G. A *vertex separator* for G will be a set of vertices S such that $G - S$ is disconnected. A *vertex coloring* is a function $c : V \rightarrow Z$, where Z is the set of integers. We say that the coloring is *proper* when $c(v) = c(w)$ implies that $(v, w) \notin E$. The neighbor set $\Gamma(v)$ of a vertex v is the set of all vertices w adjacent to

v; that is, $\Gamma(v) = \{w : (v, w) \in E\}$. A *clique* in a graph is a set of mutually adjacent vertices. A *simplicial* vertex in a graph is a vertex v such that $\Gamma(v)$ is a clique. The complete graph on n vertices, K_n, is the graph with n vertices, every two of which are adjacent.

Chordal graphs admit orderings, $v_1, v_2, \ldots, v_n$, on the vertex set such that, for each i, $N_i = \Gamma(v_i) \cap \{v_{i+1}, v_{i+2}, \ldots, v_n\}$ is a clique. These orderings are called *perfect elimination schemes.* A particular class of chordal graphs, called k-trees, is characterized by the requirement that $|N_i| = \min\{k, n-i\}$ for each i. The class of k-trees also has the following recursive definition: The complete graph on k vertices is a k-tree; if $G = (V, E)$ is a k-tree and $S \subset V$ is a k-clique, then the graph formed by adding a new vertex v and attaching it to each vertex in S is also a k-tree. Each k-tree may be constructed using several different sequences of these operations. The initial set $S \subset V$ is called a *basis* for the k-tree. $G = (V, E)$ is a *partial k-tree* if there exists a graph $G' = (V, E')$ such that $E \subseteq E'$ and G' is a k-tree. This is also expressed by saying that G has treewidth *bounded* by k. Note that, if $\sigma = v_1, v_2, \ldots, v_n$ is a perfect elimination scheme for the k-tree G, then we can construct G by first making a clique out of the vertices $\{v_{n-k+1}, v_{n-k+2}, \ldots, v_n\}$, repeatedly adding a vertex v_i, and making it adjacent to every vertex inside some k-clique.

It is easy to see that, if a graph G is chordal, then so is every vertex-induced subgraph G' of G. Thus, if a graph G can be c-triangulated, then so can every subgraph $G' \subset G$. This is expressed by saying that the property of being c-triangulatable is *hereditary.*

The only nonstandard definition that we will use frequently in this paper is the following: For a graph $G = (V, E)$ and vertex separator $S \subset V$ with C a component of $G - S$, we define $C \cup cl(S)$ to be the graph formed by adding to the subgraph of G induced by $C \cup S$ sufficient edges to make S into a clique. We also say that a graph G is a *k-partition intersection graph* if it is edge-covered by k-cliques. We may simply say that it is a partition intersection graph if the constant k is understood.

3. Lemmas and theorems.

Let G be a graph with n vertices and proper vertex coloring $c : V \rightarrow \{1, 2, \ldots, k+1\}$. We will show how we can determine in $O((n + m(k-1))^{k+2})$ time whether G can be c-triangulated, so that we will have a polynomial algorithm for the case where the number of colors is fixed. The algorithm is a modification of the partial k-tree recognition algorithm of Arnborg, Corneil, and Proskurowski [1], which determines whether an arbitrary graph has an embedding in a k-tree (see also [21] for a partial k-tree parsing algorithm based upon [1]).

We will now present two results (without proof) from our paper, which form the basis of our dynamic programming algorithm, to motivate the ensuing lemmas and theorems. The proofs of these results will follow in the text.

Let $G = (V, E)$ be a graph properly vertex-colored by c with $k+1$ colors present. Then G can be c-triangulated if and only if there exists a set $S \subset V$ of k vertices such that $G - S$ is separated, and for every component C of $G - S$, the graph $C \cup cl(S)$ can be c-triangulated.

This result appears as Lemma 4.

Let $G = (V, E)$ be a partition intersection graph properly vertex-colored by c with $k+1$ colors present. Let $S \subset V$ be a set of k distinctly colored vertices such that $G - S$ is disconnected and C is a component of $G - S$. We can determine the answer *for $C \cup cl(S)$ (i.e., whether $C \cup cl(S)$ can be c-triangulated) simply by examining the answer for every graph $C' \cup cl(S')$, where S' is a vertex separator for G consisting of k distinctly colored vertices, and C' is a component of $G - S'$ with $|V(C' \cup cl(S'))| <$*

$|V(C \cup cl(S))|$.

This result appears as Theorem 3.

The dynamic programming algorithm for partition intersection graphs is then obvious: Determine all sets S of k vertices such that $G - S$ is disconnected and determine the answers for each graph $C \cup cl(S)$, where C is a component of $G - S$, using the information computed for the smaller graphs.

However, as our input graphs may not be partition intersection graphs, we begin with the following result.

LEMMA 1. *Let $G = (V, E)$ be a graph properly colored by coloring c with $k + 1$ colors. Then there exists a partition intersection graph $G' = (V', E')$ such that the following is true*:

- *For every separator S of G' of k distinctly colored vertices and every component C of $G' - S$, $C \cup S$ has $k + 1$ colors present,*
- *G can be c-triangulated if and only if G' can be c-triangulated, and*
- *The number of vertices in G' is $n + m(k - 1)$, where $n = |V|, m = |E|$.*

Proof. For each edge $e = (v, w)$ in E, add $k - 1$ vertices and sufficient edges so that the $k + 1$ vertices together form a $(k + 1)$-clique, of distinctly colored vertices. Call the resultant graph G'. It is clear that G' is edge-covered by $(k + 1)$-cliques, and hence G' is a partition intersection graph. Therefore, for every separator S of k distinctly colored vertices and every component C of $G' - S$, $C \cup S$ has all $k + 1$ colors present. G' contains G as an induced subgraph, and it is thus easy to see that G can be c-triangulated if and only if G' can be c-triangulated. Finally, each edge in G contributes an additional $k - 1$ vertices, so that the last condition holds as well. □

Graphs that arise from inputs to the PPP are partition intersection graphs, so that, for every separator S and component C of $G - S$, $C \cup S$ have all $k + 1$ colors present. Therefore, for graphs G that arise from instances to the PPP (using Buneman's reduction in [5]), we need not make any transformation to the graph. Otherwise, we will first construct the partition intersection graph G' from G using the construction in Lemma 1 and then apply the dynamic programming algorithm to G'.

The next lemma shows that, if G can be c-triangulated, then, in fact, we can complete G to a k-tree, without ever adding edges between vertices of the same color. This allows us to use the powerful theorems about graphs of bounded treewidth to deduce structural information about G.

LEMMA 2. *Let G be a connected vertex-colored graph with vertex coloring $c : V \to \{1, 2, \ldots, k + 1\}$, with all $k + 1$ colors represented. Then G has a c-triangulation if and only if G has a c-triangulation that is a k-tree.*

Proof. One direction is obvious. Now consider the case where G has a c-triangulation. We will prove that G has a c-triangulation that is a k-tree by induction on n and k. The case where $k = 1$ implies a two-coloring on the vertices of G. It is easy to see that a connected two-colored graph G can be c-triangulated if and only if G is acyclic, i.e., is a tree. The set of 1-trees is the set of trees, and so the lemma holds whenever $k = 1$.

Let G be a colored n-vertex graph, G' a c-triangulation of G, and $c : V \to \{1, 2, \ldots, k + 1\}$ the vertex coloring. The base case, in which $n = k + 1$, is trivial. Inductively, we will assume the lemma is true for all graphs with fewer than n vertices or with fewer colors represented.

Let v be a simplicial vertex in G'. We have two cases to consider, depending on whether $G - \{v\}$ is k- or $(k+1)$-colored.

Case 1. $G - \{v\}$ is k-colored. In this case, v is the only vertex in G that is colored $c(v)$, and $G - \{v\}$ has a c-triangulation $(G' - \{v\})$. Therefore, the inductive hypothesis proves that $G - \{v\}$ has a c-triangulation $G'' = (V, E'')$, which is a $(k-1)$-tree. If we attach v to every vertex of G'', we get a graph $G''' = (V, E''')$, where $E''' = E'' \cup \{(v, w) : w \in V\}$. G''' is a k-tree, as can be seen by placing v at the end of the perfect elimination scheme for G'', and G is a subgraph of G'''. Thus, G can be c-triangulated to a k-tree.

Case 2. $G - \{v\}$ is $(k+1)$-colored. The neighbor set of v in G' will be $A = \{x_1, x_2, \ldots, x_r\}$, and, since v is simplicial in G', this is an r-clique, for some $r \leq k$.

Let G'' be a k-tree c-triangulation of $G' - \{v\}$, with perfect elimination scheme σ. Among the vertices in A, let x be the first to appear in σ. There are two cases to consider, depending on the position for x in the order given by σ. If x appears in the last $k+1$ positions, in σ, then we will connect v to each of the vertices in the last $k+1$ positions, which are colored differently from $c(v)$. This produces a k-tree, since the last $k+1$ vertices in σ are a clique in G'', and thus exactly k of these vertices are colored differently from v. We have attached v to each vertex in a k-clique, including A. Therefore, we have a k-tree supergraph of G, which is properly colored by c.

On the other hand, if x appears at least $k+2$ positions from the end, then let B denote the set of vertices that are neighbors of x in G'' and that follow x in the ordering σ. B will be a k-clique, since G'' is a k-tree, and will include $A - \{x\}$. Consider the graph G''', which is defined by attaching the vertex v to x and to every vertex in B that is colored differently from $c(v)$. Thus, $G''' = (V, E''')$, where $E''' = E(G'') \cup \{(v, y) : y \in B, c(y) \neq c(v)\}$. This graph G''' contains G as a subgraph and is a k-tree. □

We now observe certain facts about partial k-trees.

LEMMA 3. *If G can be c-triangulated to a k-tree, then any k-clique in G' can be a basis for G'.*

Proof. Let G' be a k-tree c-triangulation of G and let S be any k-clique in G'. It is shown in [15] that any k-clique of a k-tree can be a basis; hence, S can be a basis for G'. □

The following lemma has an uncolored counterpart in [1].

LEMMA 4. *A given $(k+1)$-colored graph G of size at least $k+2$ can be c-triangulated if and only if there is a set S of k distinctly colored vertices such that, for each component C of $G - S$, the graph $C \cup cl(S)$ can be c-triangulated.*

Proof. If G can be c-triangulated, then, by Lemma 2, there exists a k-tree G' that is a supergraph of G. It is shown in [16] that the minimal vertex-separators of G' are k-cliques. Let S be one such minimal vertex separator of G'. Since G' is properly colored, S consists of distinctly colored vertices, and, for each component C of $G - S$, $C \cup cl(S)$ is a subgraph of the partial k-tree given by the subgraph of G' induced by the vertices of $C \cup S$. Hence, $C \cup cl(S)$ can be c-triangulated as well.

Conversely, suppose that there is a set S of k distinctly colored vertices such that, for each component C of $G - S$, the graph $C \cup cl(S)$ can be c-triangulated. Let T_C be a c-triangulation of $C \cup cl(S)$. We can combine these graphs into a c-triangulation for all of G, since they only intersect on S. □

We have thus reduced the problem of determining whether the entire $(k+1)$-colored graph G can be c-triangulated to looking at graphs of the form $C \cup cl(S)$, where S is a a vertex separator for G of k vertices, and C is one of the components of $G - S$.

Rose, Tarjan, and Lueker [17] proved the following lemma.

LEMMA 5. *Let G be a triangulated graph, σ a perfect elimination scheme for G, and let a, b be vertices in G. If there is a path P from a to b in G such that every vertex in $P - \{a, b\}$ comes before a and b in the ordering σ, then (a, b) is an edge in G.*

We now prove the following theorem.

THEOREM 1. *Let $G = (V, E)$ be a $(k+1)$-colored partition intersection graph containing at least $k+2$ vertices, S_0 a set of k vertices of G that is a separator for G, and let C be a component of $G - S_0$. Then $C \cup cl(S_0)$ can be c-triangulated if and only if there exists a family $\mathcal{F}$ of k-trees and a vertex $v \in C$ such that*

1. *For every $F \in \mathcal{F}$, there exists a vertex $x \in S_0$ such that $V(F) = C' \cup cl(S)$ where $S = S_0 \cup \{v\} - \{x\}$, C' is both a component of $G - S$ and of $C \cup cl(S_0) - S$,*
2. *$|V(F)| < |V(C \cup cl(S_0))|$, for every $F \in \mathcal{F}$,*
3. *Every two graphs in $\mathcal{F}$ intersect only on $S_0 \cup \{v\}$, and*
4. *$G|(V - S_0)$ is contained in $\bigcup_{F \in \mathcal{F}} F$.*

Proof. It is easy to see that, if these conditions hold, we can combine the k-trees in $\mathcal{F}$ into one k-tree covering $C \cup cl(S_0)$, since they only intersect on $S_0 \cup \{v\}$. Thus, we need only show the converse.

So, suppose that $G_1 = C \cup cl(S_0)$ can be c-triangulated. Let G' be a c-triangulation of $C \cup cl(S_0)$. By Lemma 3, the k-clique S_0 can be a basis for G'. Let v be the vertex added to the basis S_0 in the construction of G' and let $S' = S_0 \cup \{v\}$. Thus, there is a perfect elimination scheme for G' in which the vertices of S' occur at the end. We will show that we can decompose $C \cup cl(S_0)$ into the union of k-trees, T_K, each of which is based upon a k-clique subset $K \subset S'$. We will then show that each such K forming the basis of one of these k-trees will be a separator for G, so that $T_K - K$ has components $C_1, \ldots, C_r$. We can then, in turn, write each T_K as the union of possibly smaller k-trees, $T_K^i = T_K|(C_i \cup K)$. These k-trees are the ones of interest.

G' is built by adding vertices, one at a time, and making each new vertex adjacent to every vertex in some k-clique. We will define G_i to be the subgraph of G' induced by the vertex set $\{v_i, v_{i+1}, \ldots, v_n\}$. Thus, G_{n-k+1} is a k-clique, and, to form G_i, we make vertex v_i adjacent to every vertex in some k-clique in G_{i+1}. We will show that we can assign to each added vertex v_i (with $i < n - k$) a label $L(v_i)$ the *name* of a k-clique $K \subset S'$, so that, for each $K \subset S'$, the subgraph $T_K = G|V_K$, where $V_K = \{v : L(v) = K \text{ or } v \in K\}$, is a k-tree. We will also show that every edge e in $G|(V - S')$ is in one of these k-trees and that the k-cliques K forming the basis of the k-trees T_K are separators of G. We will also need to show that the component C' of $C \cup cl(S_0) - L(v)$ containing v is a component of $G - L(v)$. This will prove our assertions.

We first must show how we assign vertices to k-clique subsets of S'. Let L be the assignment function we wish to define for every vertex not in S'. Suppose that we have constructed the graph G_{i+1}, are now adding v_i to the graph, and making it adjacent to every vertex in some k-clique, R. If $R \subset S'$, then we set $L(v_i) = R$. Otherwise, the vertices in R will consist of (perhaps) some unlabelled vertices (these will be in S') and at least one labelled vertex. If all of the labels in R agree, then this is the label that we will assign to v_i. On the other hand, suppose, for our construction, that when we make v_i adjacent to every vertex in the k-clique R, not all the labels are the same and that this is the first vertex in this construction for which this happens. In this case, for some vertices v_j and v_k in R, $L(v_j) = X$ and $L(v_k) = Y$, for distinct subsets $X, Y \subset S'$. Without loss of generality, we can assume that $i < j < k$. In constructing G_j, we made v_j adjacent to every vertex in some k-clique $C \subset G_{j+1}$. Note that $v_k \in C$, since v_j and v_k are adjacent and $k > j$. Since we were able to set $L(v_j) = X$ unambiguously, this means that either every vertex in C was unlabelled, and thus $X = C$, or that the labelled vertices were all labelled X. Since we have assumed that v_k was labelled, we can infer that $L(v_k) = X$, and hence $X = Y$. Thus, this assignment of vertices to k-clique bases is well defined, and each label denotes a subset K of S'. It is easy to see that the subgraph $T_K = G'|V_K$ (for $V_K = \{v : L(v) = K \text{ or } v \in K\}$) is a k-tree and that T_K is based upon the set K.

By our construction of the labelling function, it is also clear that no edge in G has different labels at its endpoints, so that every edge in $G|(V - S')$ is in exactly one k-tree, T_K.

We now show that each k-clique $K \subset S'$ forming the basis of a k-tree in $\mathcal{F}$ is a separator for $C \cup cl(S_0)$ and for G. We first show that K is a separator for $C \cup cl(S_0)$. Suppose to the contrary, so that, for some set $K \subset S$ forming the basis of a k-tree T_K, $C \cup cl(S_0) - K$ is connected. Let $K = S - \{x\}$. We will show that there is no path from x to any vertex in $C \cup cl(S_0) - K$. Let σ' be a perfect elimination scheme for $T_K \cup \{x\}$. Clearly, we can assume that x is the last vertex in σ' to occur before the vertices of K. Let a be the vertex immediately preceding x. If there is a path from x to a in $C \cup cl(S) - K$, then the edge (a, x) is in G by Lemma 5. Then, however, $S \cup \{a\}$ is a $(k + 2)$-clique, contradicting that G is a partial k-tree. The proof can be modified to show that K is a separator for G as well. Hence the k-trees T_K^i each contain fewer vertices than G.

We now complete our proof by showing that the components of $C \cup cl(S_0) - K$ are also components of $G - K$, where K is the basis of a k-tree $F \in \mathcal{F}$. Recall that, by our construction, each such basis K is a set $L(a)$ for some $a \in V(F) - K$. So let C' be a component of $C \cup cl(S_0) - L(a)$, for some $a \in C$. It is easy to see that $L(a) = S_0 \cup \{v\} - \{x\}$ is a separator for $C \cup cl(S_0)$ and that every component X of $C \cup cl(S_0) - L(a)$, such that $x \notin X$, is also a component of $G - L(a)$. Thus, we will show that $x \notin C'$, so that C' is a component of $G - L(a)$.

Suppose that $x \in C'$. Then x is adjacent to at least one vertex z of $C' - \{x\}$. When we labelled the vertex z, we labelled it with $L(a)$, implying that $x \in L(a)$, and yet, by our construction, $x \notin L(a)$. Hence, the component C' of $C \cup cl(S_0) - L(a)$ containing a is a component of $G - L(a)$. This completes our proof. □

We can now prove the following theorem, which is a colored version of a lemma in [1].

THEOREM 2. *Let $G = (V, E)$ be a $(k + 1)$-colored graph with $|V| \geq k + 2$. Let S_0 be a set of k distinctly vertices in G such that $G - S_0$ is separated and let C be a component of $G - S_0$. Then $C \cup cl(S_0)$ can be c-triangulated if and only if there exists some vertex v in C and a family of k-vertex sets $\mathcal{M}$ such that the following is true:*

1. *For each $M \in \mathcal{M}$, $M \subset S_0 \cup \{v\}$, and M is a separator for $C \cup cl(S_0)$ and for G,*
2. *For each vertex $x \in S_0 - \{v\}$, there is a $M_x \in \mathcal{M}$ and a component C_x of $G - M_x$ and of $C \cup cl(S_0) - M_x$ such that $|C_x| < |C|$ and $C_x \cup cl(M_x)$ can be c-triangulated,*
3. *Every edge in C is in exactly one C_x, given above.*

Proof. $\Rightarrow$ Suppose that $C \cup cl(S_0)$ can be c-triangulated and let G' be a c-triangulation of $C \cup cl(S_0)$. We can therefore apply Theorem 1 and deduce the existence of a vertex $v \in C$ such that the subgraph of G' induced by the vertices of $C \cup cl(S_0)$ can be written as the union of k-trees T_K based upon k-clique subsets $K \subset S' = S_0 \cup \{v\}$.

We will let $\mathcal{M}$ consist of these subsets K, which form the bases of the k-trees T_K. Theorem 1 then shows that $\mathcal{M}$ satisfies the conditions above.

$\Leftarrow$ For the converse, if such a family $\mathcal{M} = \{M_i : i \in I\}$ of vertex separators exists, then there exists $v \in C_0$ such that the graph $C \cup cl(S_0)$ is contained in the union of c-triangulatable graphs of the form $C_x \cup cl(M)$, where each $M \in \mathcal{M}$ is a k-vertex subset of $S_0 \cup \{v\}$ and C_x is a component of $G - M$ and a proper subset of C. These graphs are $(k+1)$-colored, since G is a $(k+1)$-partition intersection graph, and hence can be completed to properly colored k-trees T_x, where $V(T_x) = V(C_x \cup M)$. This family of k-trees $\mathcal{F} = \{T_x : x \in C_0 - \{v\}\}$ satisfies Theorem 1, so that $C \cup cl(S_0)$ is c-triangulatable. □

We can now conclude with our final theorem, referred to in the beginning of this section.

THEOREM 3. *Let $G = (V, E)$ be a $(k+1)$-colored partition intersection graph and let $S \subset V$ be a set of k vertices such that $G - S$ is disconnected and C is a component of $G - S$. Then we can determine whether $C \cup cl(S)$ can be c-triangulated simply by knowing the "answer" for each smaller graph of the form $C' \cup cl(S')$, where S' is a vertex separator for G of size k and C' is a component of $G - S'$.*

4. The algorithm. We first modify (if necessary) G using the techniques of Lemma 1 so that the resultant graph G' is a partition intersection graph. We can therefore assume that the input graph G is a $(k+1)$-colored partition intersection graph.

The algorithm determines which sets S of k vertices in G are vertex separators, computes all graphs $C \cup cl(S)$ for components C of $G - S$, and orders these graphs $C \cup cl(S)$ by the number of vertices. By performing a dynamic programming algorithm and computing the answers for each such graph in the order given by the number of vertices, we can efficiently compute the answer for all graphs of the form $C \cup cl(S)$, for S a separator of size k and C a component of $G - S$. This is the basis of our algorithm.

Note that the obvious recursive algorithm will not produce the running time we want, since we would have to determine many times over whether the various graphs $C \cup cl(S)$ are c-triangulatable. Here then is the dynamic programming algorithm.

INPUT: A graph G, with n vertices, and a proper vertex coloring $c : V \rightarrow \{1, 2, \ldots, k+1\}$. We assume that the number of colors present in G is $k+1$ and that G is a partition intersection graph.

OUTPUT: YES or NO.

DATA STRUCTURE: A family $\mathcal{X} = \{M_i\}$ of k-element subsets of distinctly colored vertices, which are vertex separators of G. For each set M_i in $\mathcal{X}$ and for each of the r_i components C_j of $G - M_i$, we denote by M_i^j, $j = 1, 2, \ldots, r_i$, the subgraph of G induced by $C_j \cup M_i$ with the addition of edges required to make M_i into a clique. Each such M_i^j can either be c-triangulated or not. This will be determined during the algorithm, in the order of increasing size of the M_i^j's, and an appropriate answer ("yes" or "no") will be stored for each.

```
ALGORITHM:
    for each set $M$ of $k$ distinctly colored vertices
    in $G$, do
      if $M$ is a vertex separator of $G$, then
        insert $M_i = M$ and the corresponding
            graphs $M_i^j$ into the data structure
        end-do
    sort the graphs $M_i^j$ by increasing size
    {examine the $M_i^j$ in turn by order of number of
    {vertices, and determine whether each can be
    {$c$-triangulated. Any properly colored graph of
    {$k+1$ vertices containing $k+1$ colors can be
    {$c$-triangulated ;
        if $M_i^j$ has size $k+1$, set its answer to "yes."
    {We will now apply Theorem 2 to each graph
    {$M_i^j$ and search for a vertex $v \in M_i^j - M_i$
    {and family $\mathcal{M}$ satisfying the conditions of
    {Theorem 2 to determines whether $M_i^j$ can be
    {$c$-triangulated.}
    for each graph $M_i^j$ in order of size $h > k+1$, do
        for each  $v \in M_i^j - M_i$ such that
        for all $x \in M_i$, $c(v) \neq c(x)$, do
        {We now check whether for vertex $v$ there is a
        {family $\mathcal{M}$ satisfying the
        {conditions of Theorem 2
            examine all sets $M_m$ of $k$ distinctly
                colored vertices in $M_i \cup \{v\}$
                which are vertex separators of $G$
            for each such $M_m$, let $L_m$ be the
            union of the $M_m^j$ that can be
            $c$-triangulated
            If the union of the $L_m$ (for each
            $M_m$ above) contains $M_i^j - M_i - \{v\}$,
                then set the answer of $M_i^j$ to
                "yes" and exit-do
        end-do
        if no answer was set for $M_i^j$,
            then set the answer for $M_i^j$ to "no"
        {Applying Lemma 4 now}
        if $G$ has a vertex separator $M_i$ such that
            all $M_i^j$ graphs have the answer "yes,"
            then ($G$ can be $c$-triangulated)
                return (Yes)
        else return (No)
    end-do
end of algorithm
```

Essentially, the only difference between this algorithm and the partial k-tree recognition algorithm of Arnborg, Corneil, and Proskurowski in [1] is that, for partial k-tree recognition, we check all subsets of k vertices and we check only those subsets of k vertices that are distinctly colored.

5. Analysis of running time. The implementation for this algorithm is the same as for the partial k-tree recognition algorithm of Arnborg, Corneil, and Proskurowski in [1]; the running time is therefore $O(n^{k+2})$. We describe the implementation suggested for [1].

If necessary, we will enlarge the graph $G = (V, E)$ so that it is edge-covered by $(k+1)$-cliques, using the techniques of Lemma 1. This may add $m(k-1)$ vertices to G, where m is the number of edges in G and the graph has $k+1$ colors. So let us assume that the input graph G is edge-covered by $(k+1)$-cliques and that n is the number of vertices in G.

In the worst case, the algorithm checks all subsets of k distinctly colored vertices, of which there are at most $O(n^k)$. Each of these is checked for being a vertex separator, which takes $O(n^2)$ time. The subgraphs M_i^j are bucket-sorted according to size; this can be accomplished in time $O(n^{k+1})$, and it costs only constant time to check the exit conditions for each M_i^j. Each subgraph M_i^j has at most n vertices, and we can access the vertex-separators M_m in constant time as well. Computing the union of the L_m is of the order of the size of M_i^j. Thus, the overall complexity is $O(n^{k+2})$.

Note that, if the graph G had not been edge-covered by $(k+1)$-cliques, the transformation from G to G' would increase the number of vertices to $n + m(k-1)$. Therefore, for arbitrary graphs containing n vertices, m edges, and $k+1$ colors, the complexity of this algorithm is $O((n + m(k-1))^{k+2})$.

An alternative approach to this problem would have modified the algorithm in [1] so that the list would have included all graphs of the form $C \cup cl(S)$, where S is a separator of size *up to k*. This approach avoids having to explicitly enlarge G so that it is a partition intersection graph, but would have involved a somewhat complicated case analysis in the inner loop of the algorithm.

6. Further problems. The similarities between partial k-trees and colored graphs that have c-triangulations led us to this algorithm. It is interesting to note that it is possible to use this algorithm to find a minimum c-triangulation of G, by keeping track of the minimum c-triangulation of each of the subgraphs, just as the algorithm for partial k-tree recognition achieves that for the uncolored case.

There are several more efficient algorithms for partial k-tree recognition that, unfortunately, do not carry over to this problem. The best of these is the algorithm of Bodlaender [2], which exploits the fact that partial k-trees are finite-state to obtain a running time of $O(n)$ time for each fixed k. This approach will not work for the TCGP, because Bodlaender, Fellows, and Warnow [3] showed that the class of k-colored triangulatable graphs is not finite-state. As a result, standard techniques to find efficient algorithms for recognizing these graphs will fail. Thus, an interesting open problem is whether there exists an deterministic algorithm to determine whether a given n-vertex k-colored graph can be c-triangulated with running time $O(C_k p(n,k))$, for some constant C_k depending on k, and $p(n,k)$ a polynomial in n and k.

7. Acknowledgments. The authors are grateful to the referees whose extensive comments led to tremendous improvements in the exposition. Sampath Kannan pointed out the trick (of embedding each graph in a partition intersection graph), which simplified the algorithm, and Martin Vingron helped to clarify the exposition by a careful reading and criticism of the text. The second author thanks Gene Lawler, Simon Tavare, and Michael Waterman.

REFERENCES

[1] S. Arnborg, D. Corneil, and A. Proskurowski, *Complexity of finding embeddings in a k-Tree*, SIAM J. Algebraic Discrete Meth., 8 (1987), pp. 277–284.

[2] H. Bodlaender, *A linear time algorithm for finding tree-decompositions of small treewidth*, in Proc. of the 25th Ann. Sympos. on Theory of Computing, 1993, pp. 226–234.

[3] H. BODLAENDER, M. FELLOWS, AND T. WARNOW, *Two strikes against the Perfect Phylogeny Problem*, in Proc. of ICALP, Vienna, Austria, 1992.

[4] H. BODLAENDER AND T. KLOKS, *A simple linear time algorithm for triangulating three-colored graphs*, J. Algorithms, 15 (1993), pp. 160–172.

[5] P. BUNEMAN, *A characterization of rigid circuit graphs*, Discrete Math., 9 (1974), pp. 205–212.

[6] J. CAMIN AND R. SOKAL, *A method for deducing branching sequences in phylogeny*, Evolution, 19 (1965), pp. 311–326.

[7] G. F. ESTABROOK AND L. LANDRUM, *A simple test for the possible simultaneous evolutionary divergence of two aminoacid positions*, Taxon., 24 (1975), pp. 609–613.

[8] G. F. ESTABROOK AND C. MEACHAM, *Compatibility methods in systematics*, Ann. Rev. Ecol. Syst., 16 (1985), pp. 431–446.

[9] W. M. FITCH, *Toward finding the tree of maximum parsimony*, in Proc. 8th Internat. Conf. on Numerical Taxonomy, G. F. Estabrook, ed., W. H. Freeman, San Francisco, 1975, pp. 189–230.

[10] D. GUSFIELD, *Efficient algorithms for inferring evolutionary trees*, Networks, 21 (1991), pp. 19–28.

[11] R. IDURY AND A. SCHAFFER, *Triangulating three-colored graphs in linear time and linear space*, SIAM J. Discrete Math., 6 (1993), pp. 289–293.

[12] S. KANNAN AND T. WARNOW, *Triangulating* 3*-colored graphs*, SIAM J. Discrete Math., 5 (1992), pp. 249–258.

[13] ———, *Inferring evolutionary history from* DNA *sequences*, SIAM J. Comput., 23 (1994), to appear.

[14] F. R. MCMORRIS AND C. A. MEACHAM, *Partition intersection graphs*, Ars Combinatorica, 16 (1983), pp. 135–138.

[15] A. PROSKUROWSKI, *Separating subgraphs in k-trees: Cables and caterpillars*, Discrete Math., 49 (1984), pp. 275–285.

[16] D. J. ROSE, *On simple characterization of k-trees*, Discrete Math., 7 (1974), pp. 317–322.

[17] 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.

[18] M. A. STEEL, *The complexity of reconstructing trees from qualitative characters and subtrees*, J. Classification, 9 (1992), pp. 91–116.

[19] T. J. WARNOW, *Combinatorial Algorithms for Constructing Phylogenetic Trees*, Ph. D. thesis, University of California, Berkeley, CA, May 1991.

[20] E. O. WILSON, *A consistency test for phylogenies based upon contemporaneous species*, Systematic Zoology, 14 (1965), pp. 214–220.

[21] T. WIMER, *Linear Algorithms on k-Terminal Graphs*, Ph. D. thesis, Clemson University, Clemson, SC, 1987.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 307-313, May 1994

BIPARTITE SUBGRAPHS OF TRIANGLE-FREE GRAPHS*

SVATOPLUK POLJAK† AND ZSOLT TUZA‡

Abstract. The authors present a lower bound on the maximum size of a bipartite subgraph of a triangle-free graph that improves a result due to Erdös and Lovász. It also gives a polynomial-time algorithm, while the previous bound was proved by probabilistic methods.

Key words. triangle-free graph, maximum bipartite subgraph, polynomial-time algorithm

AMS subject classification. 68R10

Erdös and Lováz (see [6]) proved by probabilistic methods that every triangle-free graph with m edges contains a bipartite subgraph with at least

$$\frac{1}{2}m + cm^{2/3}\left(\frac{\log m}{\log\log m}\right)^{1/3}$$

edges, for some positive constant c. We improve their bound to

$$\tfrac{1}{2}m + c(m\log m)^{2/3}$$

and show that a bipartite subgraph with at least that many edges can be constructed by a polynomial-time algorithm. Our approach is related to the proof of Erdös and Lovász. For a given triangle-free graph, (i) color the graph by relatively few colors and identify each color class with a single vertex of a weighted complete graph, where an edge weight corresponds to the number of edges between two color classes in G; (ii) use a symmetrization argument to establish the existence of a large cut of the complete graph obtained.

We show that both steps can be algorithmized. A suitable coloring can be obtained by successive deletion of "large" independent sets constructed by a method of Shearer, and the symmetrization argument can be turned into a construction by a trick due to [9]. The total complexity of the algorithm is bounded by $O(n^{11/3}\log^{-2/3} n)$. All logarithms are of base e.

LEMMA 1 (Shearer [13]). *Let G be triangle-free graph with n vertices and average degree d. Let $f(d) = (d\log d - d + 1)/(d-1)^2, f(0) = 1, f(1) = 1/2$. Then G contains an independent set of at least $nf(d)$ vertices.*

As proved in [13], an independent set S of required size can be obtained by the following procedure. Let $f'(d)$ be the derivative of the above function f. Find a vertex u (it must exist) satisfying

$$(d_1+1)f(d) \leq 1 + (dd_1 + d - 2d_1d_2)f'(d),$$

where d_1 and d_2 denote the degree of u and the average degree of the neighbours of u, respectively. Set $S := S' \cup \{u\}$, where S' is the independent set constructed

* Received by the editors March 18, 1991; accepted for publication (in revised form) February 4, 1993.

† Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí25, 118 00 Praha 1, Czech Republic. The author is on leave at the Institute für Diskrete Mathematik, Universität Bonn. This author's research was supported by the A. von Humboldt Foundation.

‡ Computer and Automation Institute, Hungarian Academy of Sciences, H-1111 Budapest, Kende u.13-17, Hungary.

recursively in the subgraph G' after deleting u and the set $N(u)$ of its neighbours from G. Clearly, vertex u can be found in $O(n^2)$ time, and hence the complexity of constructing S is $O(n^3)$.

We also use Lemma 1 in a more convenient form: A triangle-free graph G with n vertices and average degree $d \geq 2$ contains an independent set of at least $\frac{3}{4}n(\log d/d)$ vertices.

LEMMA 2. *A triangle-free graph G with $n \geq 15$ vertices contains an independent set of size at least $(\frac{3}{8}n \log n)^{1/2}$, which can be constructed in $O(n^3)$ time.*

Proof. If G contains a vertex u of degree greater or equal to $d' := (\frac{3}{8}n \log n)^{1/2}$, then the neighbours of u form an independent set of the required size. Assume there is no such vertex. Then the average degree d is less than d'. Assume $d \geq e$. By the corollary of Lemma 1, we have

$$\alpha(G) \geq \frac{3}{4}n\frac{\log d}{d} \geq \frac{3}{4}n\frac{\log d'}{d'} \geq \frac{3}{4}n\frac{\frac{1}{2}(\log \frac{3}{8} + \log n + \log\log n))}{(\frac{3}{8}n \log n)^{1/2}}.$$

Since $\log \frac{3}{8} + \log\log n > 0$ for $n \geq 15$, we have

$$\alpha(G) \geq \frac{3}{4}n\frac{\frac{1}{2}\log n}{(\frac{3}{8}n \log n)^{1/2}} = \left(\frac{3}{8}n \log n\right)^{1/2}.$$

If $d \leq e$, then

$$\alpha(G) \geq nf(d) \geq nf(e) \geq \left(\frac{3}{8}n \log n\right)^{1/2}$$

by Lemma 1 and because the function $f(d)$ is decreasing. □

THEOREM 1. *The chromatic number $\chi(G)$ of every triangle-free graph G of n vertices and m edges is bounded by*

$$\min\left\{4\left(\frac{n}{\log n}\right)^{1/2}, 14\frac{m^{1/3}}{(\log m)^{2/3}}\right\},$$

and proper vertex colorings with that many colors can be found in $O(n^{11/3}\log^{-2/3} n)$ time.

Proof. (i) Define $f(x) = 4(x/\log x)^{1/2}$. We show that $\chi(G) \leq f(n)$. Since the general induction step below applies only for sufficiently large value of n, some initial values should be handled separately. For this purpose, observe a simple bound $\chi(G) \leq \lfloor n/2 \rfloor + 1$. (Choose a pair x, y of nonadjacent vertices and color them by the first color. Since G is triangle-free, $G\backslash\{x, y\}$ also contains nonadjacent pair, and so on.) Since $\lfloor n/2 \rfloor + 1 \leq f(n)$ for $n \leq 17$, we are done for these values of n. Furthermore, we compute that

$$\lfloor f(n) \rfloor = \begin{cases} 9 & \text{for } n = 18, \\ 10 & \text{for } n = 19, \ldots, 24, \\ 11 & \text{for } n = 25, \ldots, 30. \end{cases}$$

For $n = 18$, 24, and 30, the graph G contains an independent set of size $\lceil(\frac{3}{8}n \log n)^{1/2}\rceil = 5, 6$, and 7, respectively. Thus the statement is proved for $n \leq 30$. Furthermore, we will proceed by induction on n. Let $n > 30$ and assume that the statement is true for all values less than n. By Lemma 2, there is an independent set S of size at least $(\frac{3}{8}n\log n)^{1/2}$. Let $G' := G\backslash S$. By the induction hypothesis, $\chi(G') \leq f(n')$, where n' denotes the number of vertices of G'. Then $\chi(G) \leq 1 + \chi(G') \leq 1 + f(n')$. We prove $f(n') + 1 \leq f(n)$. By the "mean value theorem," we have $f(n) - f(n') = (n-n')f'(z)$ for some z, $n' \leq z \leq n$. We can check that $f'(x)$ is decreasing for $x \geq 6$. Then

$$\begin{aligned} f(n) - f(n') &= (n-n')f'(z) \geq (n-n')f'(n) \\ &\geq \left(\frac{3}{8}n\log n\right)^{1/2} 2\left(\frac{n}{\log n}\right)^{-1/2} \frac{\log n - 1}{\log^2 n} = 2\left(\frac{3}{8}\right)^{1/2} \frac{\log n - 1}{\log n}, \end{aligned}$$

which is greater than 1 for $n > 30$. Hence $\chi(G) \leq f(n)$ is proved.

(ii) Define $n^* = m^{2/3}/(\log m)^{1/3}$. If $n \leq n^*$, then we can apply (i) as follows:

$$\chi(G) \leq 4(n/\log n)^{1/2} \leq 4(n^*/\log n^*)^{1/2} \leq 4\left(\frac{m^{2/3}}{(\log m)^{1/3}} \cdot \frac{1}{\frac{2}{3}\log m}\right)^{1/2} \leq 6\frac{m^{1/3}}{(\log m)^{2/3}}.$$

Suppose that $n > n^*$. We can also eliminate vertices of small degree (either 0 or 1), i.e., assume $m \geq n$. Applying Lemma 1, we find a sequence $S_1, S_2, \ldots, S_k$ of relatively large independent sets, below:

$$|S_i| \geq \frac{3}{4}|V(G_{i-1})|\frac{\log \bar{d}(G_{i-1})}{\bar{d}(G_{i-1})},$$

where $G_0 := G, G_i := G_{i-1}\backslash S_i$, and $\bar{d}$ denotes average degree. Putting $n_i := |V(G_{i-1})|$, the procedure stops when $n_k \leq n^*$. Since

$$\chi(G_k) \leq 4\left(\frac{n^*}{\log n^*}\right)^{1/2} \leq 6\frac{m^{1/3}}{(\log m)^{2/3}},$$

it is sufficient to prove that $k \leq 6(m^{1/3}/\log m^{2/3})$. Observe that $\bar{d}(G_i) \leq 2m/n_i$, i.e.,

$$|S_{i+1}| \geq \frac{3}{4}\frac{n_i \log \bar{d}(G_i)}{\bar{d}(G_i)} \geq \frac{3}{4}\cdot\frac{n_i^2}{2m}\log\frac{m}{n_i}.$$

For $x \in S_i$, let $w(x) = |S_i|^{-1}$. Then the number k of colors is equal to $k = \sum w(x)$. The previous observation shows

$$w(x) \leq \frac{4}{3}\frac{2j^2 m}{n^2 \log\frac{jm}{n}}$$

for $x \in S_i$ when $n_{i-1} \geq n/j$. Thus, for the vertices between $n/(j-1)$ and n/j, the total weight is at most

$$\frac{4}{3}\left(\frac{n}{j-1} - \frac{n}{j}\right)^2 \cdot \frac{j^2}{n} \cdot \frac{m}{n\log\frac{jm}{n}} = \frac{4}{3}\cdot\frac{2m}{n}\cdot\frac{1}{\log j + \log\frac{m}{n}}. \tag{1}$$

Here j was from 2 to $n/n^* = n(\log m)^{1/3}/m^{2/3} := t$. If $t \leq m/n$, then we delete the first term of the denominator of (1) and obtain

$$k = \sum w(x) \leq \frac{4}{3}\cdot\frac{m}{n}\cdot\frac{n}{m^{2/3}}\frac{(\log m)^{1/3}}{\log\frac{m}{n}} = \frac{4}{3}m^{1/3}\frac{(\log m)^{1/3}}{\log\frac{m}{n}}.$$

Since we assumed $t \leq m/n$, we have $tm/n = (m \log m)^{1/3}$, i.e., $m/n > m^{1/6}$ and $\log m/n > \frac{1}{6} \log m$. Thus, $k \leq 8m^{1/3}/(\log m)^{2/3}$.

If $t \geq m/n$, then $t > m^{1/6}$, and $\log t \geq \frac{1}{6} \log m$. Deleting the second term of the denominator in (1), we obtain

$$k = \sum w(x) \leq \frac{4}{3} \cdot \frac{m}{n} \sum_{j=2}^{t} \frac{1}{\log j}.$$

Note that $\int_2^t (1/\log x)dx \leq t/\log t$. Consequently,

$$k \leq \frac{4m}{3n} \cdot \frac{t}{\log t} = \frac{4m}{3n} \cdot \frac{n}{m^{2/3}} \cdot \frac{(\log m)^{1/3}}{\log t}.$$

Since $\log t \geq \frac{1}{6} \log m$, we conclude that $k \leq 8(m^{1/3}/(\log m)^{2/3})$.

The complexity bound $O(n^{11/3} \log^{-2/3} n)$ is obtained, since we must repeat $O(m^{1/3} \log^{-2/3} m)$ times the search for a large independent set, and each search needs $O(n^3)$ time. □

Let K_r be the complete graph on r vertices and let $w(i,j)$ be a weight of the edge ij for each pair of the vertices $i, j \in \{1, \ldots, r\}$. For $S \subset V, \delta S$ denotes the set of edges between S and $V \backslash S$, and $w(\delta S) := \sum_{e \in \delta S} w(e)$ is the weight of the cut δS. *The maximum cut problem* is to find a cut δS of maximum weight $w(\delta S)$. Clearly, the maximum bipartite subgraph problem is a special case of the max-cut problem, when all edge weights are 0 or 1. The maximum bipartite subgraph problem, and hence also the maximum cut problem, are known to be NP-complete (see [8]).

LEMMA 3. *Let K_r be a complete graph, w be an edge-weight function, and $M := \sum w(e)$ be the total sum of edge weights. Then*

(i) *There exists a cut δS whose weight is at least $(\frac{1}{2} + 1/2r)M$,*

(ii) *We can construct a cut of weight at least $(\frac{1}{2} + 3/7r)M$ in time $O(r^3)$.*

Proof. (i) The following probabilistic proof is well known (see, e.g., [5]). Assume that $r = 2k$ (or $r = 2k - 1$) and consider a random 2-partition of the vertex set into two parts of sizes k (or k and $k - 1$). The probability that the endvertices of an edge are in distinct classes is $k/(2k - 1)$, which is at least $(r + 1)/2r$. Hence the solution of the max-cut problem has value at least $M(r + 1)/2r$.

(ii) An efficient construction is possible by a trick due to Lieberherr and Specker (used in [9] for a related problem of partial 2-satisfiability). The construction needs $O(r^3)$ steps and ensures the lower bound M in the case where r is a power of a prime. Otherwise, we must take some $s > r$, which is a power of a prime, embed K_r into K_s (so that the weights of new edges are zero), and find a cut of weight at least M in K_s. By a result of Breusch [1], for every integer $r \geq 48$, there exist a prime p such that $r \leq p \leq \frac{9}{8}r$. Hence there is always some prime power s with $r \leq s \leq \frac{7}{6}r$ (we must check only values $r \leq 47$). Hence we proved the bound of (ii). In fact, the gap between consecutive primes p and p' is $o(p)$ as $p \to \infty$. The construction of [9] goes as follows.

A permutation group $\mathcal{G}$ is said to be *doubly transitive* when, for every quadruple $i \neq i'$ and $j \neq j'$, there is a permutation π such that $\pi(i) = j$ and $\pi(i') = j'$. For example, the full permutation group is doubly transitive, but fortunately there are also much smaller doubly transitive groups. Let $\mathcal{G}$ be a doubly transitive group on $\{1, \ldots, s\}$. Consider the system of 2-partitions $(S_\pi, V \backslash S_\pi), \pi \in \mathcal{G}$, where $S_\pi = \{\pi(1), \ldots, \pi(\lfloor s/2 \rfloor)\}$, and observe that every edge ij of the graph G is separated by the same number of partitions $(S_\pi, V\backslash S_\pi)$. Hence, the above probabilistic proof by Erdös can be modified so that only the partitions $(S_\pi, V\backslash S_\pi)$ are considered. Thus, we have $w(\delta(S_\pi)) \geq M(s+1)/2s$ for some $\pi \in \mathcal{G}$. Hence we can check all the partitions $(S_\pi, V\backslash S_\pi), \pi \in \mathcal{G}$, and one of them is a required solution. This gives a polynomial-time algorithm, provided that we have a doubly transitive group of polynomial size. A doubly transitive group of size $s(s-1)$ can be obtained from a finite field $GF[p^\alpha], s = p^\alpha$ by taking $\mathcal{G}$ as the permutations $i \mapsto iq + r, q, r \in GF[p^\alpha], q \neq 0$. □

As a consequence of Theorem 1 and Lemma 3(i), we have the following lower bound.

THEOREM 2. *Let G be a triangle-free graph with n vertices and m edges. Then G contains a bipartite subgraph with at least*

$$\frac{m}{2} + \frac{1}{24}(m\log m)^{2/3} \tag{2}$$

edges.

For m large $(m > cn^{3/2}(\log n)^{1/2})$, a better estimate

$$\frac{m}{2} + \frac{1}{8}m\left(\frac{\log n}{n}\right)^{1/2} \tag{3}$$

can be obtained applying the first bound of Theorem 1.

CORLLARY 1. *There is an $O(n^{11/3}\log^{-2/3} n)$ time algorithm that, for a given triangle-free graph with n vertices and m edges, constructs a bipartite subgraph with at least*

$$\frac{m}{2} + \frac{1}{28}(m\log m)^{2/3}$$

or

$$\frac{m}{2} + \frac{1}{10}m\left(\frac{\log n}{n}\right)^{1/2}$$

edges.

Proof. The algorithm is as follows. We first construct an r-coloring of G with $r \leq 12m^{1/3}/(\log m)^{2/3}$ in $O(n^{11/3}\log^{-2/3} n)$ time. Then we apply the algorithm of Lemma 3(ii) to find a large cut of the complete graph K_r, where the weight $w(ij)$ is defined as the number of edges between the color classes i and j in G. The construction of the cut requires $O(r^3) = O(m/\log^2 m)$ time, which is less than the time we need to color G. □

In [7] there are two other lower bounds

$$\frac{m}{2} + \frac{2m(2m^2 - n^3)}{n^2(n^2 - 2m)} \tag{4}$$

and

$$\frac{4m^2}{n^2} \tag{5}$$

on the size of a maximum bipartite subgraph of a triangle-free graph. Let us compare bounds (2)–(5). Bound (5) is best for m greater than $\frac{1}{6}n^2$, (4) is best for m between

$n^{7/4}(\log n)^{1/4}$ and $\frac{1}{6}n^2$, (3) is best for m between $n^{3/2}(\log n)^{1/2}$ and $n^{7/4}(\log n)^{1/4}$, and (2) is best for m less than $n^{3/2}\log^{1/2} n$.

PROPOSITION 1. *Both bounds* (4) *and* (5) *can be achieved by a polynomial-time algorithm.*

Proof (sketch). In the case of (4), find an edge uv of G that is contained in the maximum number of quadrilaterals (cycles C_4). Set $S_u := N(v)$ and $S_v := N(u)$ (where $N(x)$ denotes the set of the neighbours of x), and add successively the remaining vertices $x \in V \backslash (S_u \cup S_v)$ to one of S_u and S_v, depending on where x has fewer neighbours. The resulting partition $S_u \cup S_v = V$ induces a bipartite subgraph with at least (4) edges.

The procedure for (5) is even simpler; the required bipartite subgraph is induced by a bipartition $(N(u), V\backslash N(u))$ for a suitable vertex $u \in V$. □

Even better bounds are known for some special cases. Bondy and Locke [2] proved that every triangle-free graph with degree at most 3 contains a bipartite subgraph with at least $\frac{4}{5}m$ edges, and the subgraph can be constructed in polynomial time. Locke [10] proved that a $2k$-regular triangle free graph contains a bipartite graph with $m(k+2)/(2k+2)$ edges (and a similar bound for $2k+1$-regular graphs). Some of Locke's estimates have been generalized by Caro and the second author in [3]. Zýka [15] presented a lower bound for 3-regular graphs with large girth.

In the general case, for a connected graph with possible triangles, there is a bipartite subgraph with at least $\frac{1}{2}m + \frac{1}{4}(n-1)$ edges, as proved by Edwards [4]. A polynomial-time algorithm to find such a subgraph has been given in [11] and generalized to the weighted case in [12].

Note added in proof. Recently, Shearer [14] proved that every triangle-free graph with m edges contains a bipartite subgraph with at least $m/2 + cm^{3/4}$ edges. This improves the bound of Corollary 1 when m is less than $n^2/\log^2 n$. However, it remains an open question whether the improved bound can be attained by a polynomial-time algorithm.

REFERENCES

[1] R. BREUSCH, *Zur Verallgemeinerung des Bertrandschen Postulates, dass zwischen x und $2x$ stets Primzahlen liegen*, Math. Z., 34 (1931), pp. 505–526.

[2] J. A. BONDY AND S. C. LOCKE, *Largest bipartite subgraphs in triangle-free graphs with maximum degree three*, J. Graph Theory, 10 (1986), pp. 477–504.

[3] Y.CARO AND ZS. TUZA, *Improved lower bounds on k-independence*, J. Graph Theory, 15 (1991), pp. 94–107.

[4] C. S. EDWARDS, *Some extremal properties of bipartite graphs*, Canad. Math. J., 25 (1973), pp. 475–485.

[5] P. ERDÖS, *On bipartite subgraphs of a graph*, Math. Lapok, 18 (1967), pp. 283–288.

[6] ———, *Problems and results in graph theory and combinatorial analysis*, in Graph Theory and Related Topics, J. A. Bondy and U. S. R. Murty, eds., Proc. Conf. Waterloo, 1977, Academic Press, New York, 1979, pp. 153–163.

[7] P. ERDÖS, R. FAUDREE, J. PACH, AND J. SPENCER, *How to make a graph bipartite*, J. Combin. Theory Ser. B, 45 (1988), pp. 86–98.

[8] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[9] K. J. LIEBERHERR AND E. SPECKER, *Complexity of partial satisfaction*, J. Assoc. Comput. Mach., 28 (1981), pp. 411–421.

[10] S. C. LOCKE, *A note on bipartite subgraphs of triangle-free graphs*, J. Graph Theory, 14 (1990), pp. 181–185.

[11] S. POLJAK AND D. TURZÍK, *A polynomial algorithm for constructing a large bipartite subgraph with an application to satisfiability problem*, Canad. Math. J., 24 (1982), pp. 519–524.

[12] ———, *A polynomial heuristic for certain subgraph optimization problems with guaranteed lower bound*, Discrete Math., 58 (1986), pp. 99–104.

[13] J. B. SHEARER, *A note on the independence number of triangle-free graphs*, Discrete Math., 46 (1983), pp. 83–87.

[14] ———, *A note on bipartite subgraphs of triangle-free graphs*, Random Structures Algorithms, 3 (1992), pp. 223–226.

[15] O. ZÝKA, *On bipartite density of regular graphs with large girth*, J. Graph Theory, 14 (1990), pp. 631–634.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 314–324, May 1994

SUMS OF SQUARES OF EDGE LENGTHS AND SPACEFILLING CURVE HEURISTICS FOR THE TRAVELING SALESMAN PROBLEM*

JUN GAO† AND J. MICHAEL STEELE‡

Abstract. The sum of squares of the edge lengths of the tour provided by the spacefilling curve heuristic applied to a random sample of n points from the unit square is proved to be asymptotically equal to a periodic function of the logarithm of the sample size.

Key words. spacefilling curves, traveling salesman problem (TSP), minimal spanning tree, combinatorial optimization, order statistics, sums of squares

AMS subject classifications. primary 05C05, secondary 60F15

1. Two sources of motivation.

Two lines of investigation come together to form the motivation for the present work. The first of these concerns the behavior of the sums of squares of the edge lengths in several classical problems of geometric combinatorial optimization. The second concerns recent progress in the understanding of the behavior of the spacefilling curve heuristic for the traveling salesman problem (TSP).

Sums of squares of edges. This line begins with an empirical discovery of R. Bland. Since we obtain the same minimum spanning tree (MST) for a set of n points $\{x_1, x_2, \dots, x_n\} \subset [0,1]^2$ whether we assign edge costs c_{ij} that are equal to the Euclidean length $||x_i - x_j||$ or to the square of the lengths $||x_i - x_j||^2$, we can save some computation time by working with the squared lengths. When Bland used the sum of squared edge lengths as the feature of merit in a study of algorithms for the MST, he found after computing the MST of a number of random samples of different sizes from $[0,1]^2$ that the value of the minimum value of the sum of squared edge lengths showed very little dependence on n and little variation between samples. Bland was led to conjecture that there is a constant C_{MST} such that for the MST of $\{X_1, X_2, \dots, X_n\}$ where the X_i are independent random variables with the uniform distribution on $[0,1]^2$, we have

$$\lim_{n\to\infty} \sum_{e\in\text{MST}} ||e||^2 = C_{\text{MST}} \tag{1}$$

with convergence in probability. This conjecture was proved in Aldous and Steele [1].

The method used to prove (1) relied on the possibility of calculating the MST via a greedy algorithm. Still, there are many functionals that are closely related to the MST for which there is no such possibility. Probably the most studied of these is the TSP that asks for the shortest tour through the points $\{x_1, x_2, \dots, x_n\} \subset [0,1]^2$. We do not know at present whether the analogue to (1) holds for the TSP.

For the worst-case analysis, the state of knowledge for the TSP is more complete. Snyder and Steele [10] showed that there is a universal constant C_{TSP} such that, for

* Received by the editors November 23, 1992; accepted for publication (in revised form) March 2, 1993.

† Portfolio Management International, Swiss Bank Corporation, 10 East 50th Street 32nd floor, New York, New York 10022.

‡ Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104 (steele@wharton.upenn.edu). This author's research was supported in part by National Science Foundation grant DMS92-11634 and Army Research Office grant DAAL03-91-G-0110.

any $S = \{x_1, x_2, \dots, x_n\} \subset [0,1]^2$ and any tour T of S of minimal length, we have

$$\sum_{e \in T} ||e||^2 \le C_{\mathrm{TSP}} \log n. \tag{2}$$

In subsequent analyses, Bern and Epstein [3] showed that the logarithmic term of (2) could not be replaced with a more slowly growing function.

Limit theory for the spacefilling heuristic. The spacefilling curve heuristic rests on the existence of a surjective mapping $\psi : [0,1] \to [0,1]^2$ such that for each $x \in [0,1]^2$ we can *quickly compute* a $t \in [0,1]$ such that $\psi(t) = x$. Formally, given $\{x_1, x_2, \dots, x_n\} \subset [0,1]^2$, we have a three-step process where we (1) compute a set of points $\{t_1, t_2, \dots, t_n\} \subset [0,1]$ such that $\psi(t_i) = x_i$ for each $1 \le i \le n$, (2) order the t_i so that $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$, and, finally, (3) define a permutation $\sigma : [1,n] \to [1,n]$ by requiring $x_{\sigma(i)} = \psi(t_{(i)})$. The path that visits $\{x_1, x_2, \dots, x_n\}$ in the order of $x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}$ will be called the *spacefilling curve path*, and the tour that closes this path by adding the step from $x_{\sigma(n)}$ back to $x_{\sigma(1)}$ will be called the *spacefilling curve tour.*

Here we will focus on the behavior of the spacefilling curve heuristic in the context of the simplest possible stochastic model, where the points to be toured are modeled by independent random variables X_i, $1 \le i \le n$ that are uniformly distributed in $[0,1]^2$. In [4], results of Platzman and Bartholdi [8] were refined to show that for a *large class* of spacefilling curves (SFCs) the length $L_n^{\mathrm{SFC}} = L^{\mathrm{SFC}}(X_1, X_2, \dots, X_n)$ of a spacefilling heuristic tour through $\{X_1, X_2, \dots, X_n\}$ satisfies

$$\lim_{n \to \infty} \frac{EL_n^{\mathrm{SFC}}}{\sqrt{n}\varphi(\log_p n)} = 1, \tag{3}$$

where p is an integer depending on the geometry of the spacefilling curve and where φ is a continuous periodic function of period 1 that is bounded away from zero. This behavior offers a novel contrast to that of the length $L_n^{\mathrm{OPT}} = L_n^{\mathrm{OPT}}(X_1, X_2, \dots, X_n)$ of the *shortest tour* through the random sample $\{X_1, X_2, \dots, X_n\}$, where the theorem of Beardwood, Halton, and Hammersley [2] declares that for the optimal solution no periodic term is needed; rather, there is simply a constant $\beta > 0$ such that for $n \to \infty$ we have

$$L_n^{\mathrm{OPT}} / \sqrt{n} \to \beta,$$

where the convergence takes place in expectation as well as with probability 1.

The purpose of this article is to provide a precise asymptotic understanding of the sum of the squares of the edges of the tour provided by the spacefilling curve heuristic

$$S_n = \sum_{i=1}^{n} ||X_{\sigma(i)} - X_{\sigma(i+1)}||^2 = \sum_{i=1}^{n} ||\psi(t_{(i)}) - \psi(t_{(i+1)})||^2, \tag{4}$$

where we invoke the convention that

$$\sigma(n+1) \overset{\mathrm{def}}{=} \sigma(1) \quad \text{and} \quad t_{(n+1)} \overset{\mathrm{def}}{=} t_{(1)}.$$

We will establish the possibly surprising fact that for a large class of spacefilling curves the value of this random variable is well approximated by a periodic function of the logarithm of the sample size.

2. Main result. The first three properties described below are found in many classical spacefilling curves, including those of Hilbert [5] and Peano [7]. For these and related curves, the ordinary Lipschitz property and the bimeasure-preserving property are established in Milne [6]. The dilation and translation properties are easily verified by direct consideration of the traditional constructions, and only minor alterations of the usual constructions are needed to obtain curves with the circular Lipschitz property Assumption 4.

Assumption 1 (dilation property). There is an integer $p \geq 2$ such that, for all $0 \leq s, t \leq 1$,

$$||\psi(s) - \psi(t)|| = \sqrt{p}\left\|\psi\left(\frac{s}{p}\right) - \psi\left(\frac{t}{p}\right)\right\|.$$

Assumption 2 (translation property). For $1 \leq i \leq p$, if $(i-1)/p \leq s, t \leq i/p$, then

$$||\psi(s) - \psi(t)|| = ||\psi(s + 1/p) - \psi(t + 1/p)||.$$

Assumption 3 (bimeasure-preserving property) Given any Borel set A in $[0,1]$, the set $\psi(A)$ is measurable, and

$$\lambda_1(A) = \lambda_2(\psi(A)),$$

where λ_d is the Lebesgue measure on $\mathbb{R}^d$.

Assumption 4 (circular Lipschitz property). There is a constant c_ψ such that

$$||\psi(s) - \psi(t)|| \leq c_\psi \rho(s,t)^{1/2},$$

where ρ is the circular metric on $[0,1]$ given by

$$\rho(s,t) = \min\{|s-t|, 1 - |s-t|\}.$$

The main result of this article is the following theorem.

THEOREM 1. *If a heuristic tour is built using a spacefilling curve ψ that satisfies Assumptions* 1–4, *then there exists a strictly positive continuous function φ of period* 1 *such that*

$$\lim_{n\to\infty} \frac{ES_n}{\varphi(\log_p n)} = 1, \tag{5}$$

where p is the integer of Assumptions 1 *and* 2.

3. Convergence of expectations. We first recall that a ψ that satisfies Assumption 3 creates a natural correspondence between random variables with the uniform distribution on $[0,1]^2$ and $[0,1]$. We safely omit the routine proof.

LEMMA 1. *Suppose that X is a random variable that is uniformly distributed in $[0,1]^2$ and that $\psi : [0,1] \to [0,1]^2$ is a surjection. Let ψ^* be a function that, for every $x \in [0,1]^2$, selects a preimage of x; that is, ψ^* satisfies $\psi(\psi^*(x)) = x$. For t defined by $t = \psi^*(X)$, we have the fact that t is uniformly distributed in $[0,1]$, provided that the spacefilling curve ψ satisfies the bimeasure preserving Assumption* 3.

One key to the analysis of S_n is that ES_n has a tidy expression in terms of the independent (unordered) t_i's. If $d(s,t)$ is given by

$$d(s,t) = \begin{cases} t - s & \text{if } 0 \leq s \leq t \leq 1, \\ 1 - s + t & \text{if } 0 \leq t \leq s \leq 1, \end{cases}$$

then $d(s,t)$ describes the distance along the circle of unit circumference in the counterclockwise direction from s to t, and we can write S_n in the symmetrical form

$$S_n = \sum_{i=1}^{n} ||\psi(t_i) - \psi((t_i + \delta_i) \bmod 1)||^2,$$

where

$$\delta_i = \min_{t_j \in S_i} d(t_i, t_j)$$

and $S_i = \{t_1, t_2, \dots, t_{i-1}, t_{i+1}, \dots, t_n\}$. Moreover, the variables δ_i and t_i are independent, and for each i the variable δ_i has probability density given by

$$f(t) = (n-1)(1-t)^{n-2};$$

so we can compute

$$\begin{aligned} ES_n &= E\sum_{i=1}^{n} ||\psi(t_i) - \psi((t_i + \delta_i) \bmod 1)||^2 \\ &= nE||\psi(t_1) - \psi((t_1 + \delta_1) \bmod 1)||^2 \\ &= n\int_0^1 \int_0^1 ||\psi(s) - \psi((s+t) \bmod 1)||^2 (n-1)(1-t)^{n-2} ds dt. \end{aligned}$$

Finally, introducing

$$m(t) = \int_0^1 ||\psi(s) - \psi((s+t) \bmod 1)||^2 ds,$$

we end up with the following lemma.

LEMMA 2 (key representation). *It holds that*

$$ES_n = n(n-1)\int_0^1 m(t)(1-t)^{n-2} dt. \tag{6}$$

To use this representation, we must collect some properties of $m(t)$. From the definition of $m(\cdot)$ and the circular Lipschitz property (Assumption 4) of ψ, we immediately find a useful pointwise bound.

LEMMA 3. *For $0 \le t \le 1$, we have*

$$m(t) \le c_\psi^2 \min(t, 1-t),$$

where c_ψ is the Lipschitz constant of Assumption 4.

To get to the deeper properties of $m(t)$, we first set $q = 1/p$ and define a sequence of related functions $\{f_n\}_{n\ge 0}$ on the increasing intervals $[0, p^n q]$ by

$$f_0(t) = \int_0^{1-t} ||\psi(s+t) - \psi(s)||^2 ds, \qquad 0 \le t \le q$$

and

$$f_n(t) = p^n f_0\left(\frac{t}{p^n}\right), \qquad 0 \le t \le p^n q.$$

We also define a parallel sequence of functions $\{g_n\}_{n\geq 0}$ by

$$g_0(t) = f_0(t) + c_\psi^2 t^2, \qquad 0 \leq t \leq q$$

and

$$g_n(t) = p^n g_0\left(\frac{t}{p^n}\right), \qquad 0 \leq t \leq p^n q.$$

The benefit of introducing f_n and g_n is that we can show that they share a common limit that offers insight into the behavior of $m(x)$. We first establish a monotonicity relationship.

LEMMA 4. *There is a function $w(t)$ such that for all $n \geq 0$ we have*

$$f_n(t) \leq f_{n+1}(t) \leq w(t) \leq g_{n+1}(t) \leq g_n(t), \qquad 0 \leq t \leq p^n q. \tag{7}$$

Proof. For any $0 \leq x \leq q$, we see from the definition of $m(x)$ that

$$m(x) \geq \sum_{i=1}^{p} \int_{(i-1)q}^{iq-x} ||\psi(s) - \psi(s+x)||^2 ds. \tag{8}$$

Next, by the translation Assumption 2, by changing variables, and by the dilation Assumption 1, we obtain

$$\begin{aligned} \sum_{i=1}^{p} \int_{(i-1)q}^{iq-x} ||\psi(s) - \psi(s+x)||^2 ds &= p \int_0^{q-x} ||\psi(s) - \psi(s+x)||^2 ds \\ &= \int_0^{pq-px} \left\|\psi\left(\frac{s}{p}\right) - \psi\left(\frac{s+px}{p}\right)\right\|^2 ds \\ &= \frac{1}{p} \int_0^{1-px} ||\psi(s) - \psi(s+px)||^2 ds \\ &= \frac{1}{p} m(px). \end{aligned} \tag{9}$$

By combining (9) and (8), we find the basic fact that

$$m(x) \geq \frac{1}{p} m(px), \tag{10}$$

and from (10) the first inequality of (7) follows immediately.

Turning to the g_n, we first note by the definition of g_0 and identity (9) that for $0 \leq x \leq q$ we have

(11)

$$\begin{aligned} g_0(x) &= \left\{ \sum_{i=1}^{p} \int_{(i-1)q}^{iq-x} ||\psi(s) - \psi(s+x)||^2 ds + \sum_{i=1}^{p-1} \int_{iq-x}^{iq} ||\psi(s) - \psi(s+x)||^2 ds \right\} + c_\psi^2 x^2 \\ &= \frac{1}{p} m(px) + \sum_{i=1}^{p-1} \int_{iq-x}^{iq} ||\psi(s) - \psi(s+x)||^2 ds + \left\{ c_\psi^2 p x^2 - c_\psi^2 (p-1) x^2 \right\} \\ &= \frac{1}{p} m(px) + c_\psi^2 p x^2 + \sum_{1=1}^{p-1} \left\{ \int_{iq-x}^{iq} ||\psi(s) - \psi(s+x)||^2 ds - c_\psi x^2 \right\}. \end{aligned}$$

By the Lipschitz property of ψ, we have

$$\int_{iq-x}^{iq} ||\psi(s) - \psi(s+x)||^2 ds - c_\psi^2 x^2 \leq 0;$$

so, after replacing x by x/p, we have for $0 \leq x \leq 1$ that $g_0(x) \geq pg_0(x/p)$, and our inequality for g_0 immediately yields the last inequality of (7).

Local boundedness and monotonicity of the sequences $f_n(x)$ and $g_n(x)$ tell us the sequences have pointwise limits. The definition of g_n further tells us that

$$g_n(x) - f_n(x) = c_\psi^2 x^2/p^n, \qquad 0 \leq x \leq p^n q; \tag{12}$$

so, in fact, both $f_n(x)$ and $g_n(x)$ must have the same limit; moreover, if we denote this limit by $w(x)$, then by the first and last inequalities of (7) we have for $0 \leq x \leq p^n q$ that

$$f_n(x) \leq w(x) \leq g_n(x), \tag{13}$$

completing the proof of Lemma 4.

The next lemma shows how $w(x)$ approximates $m(x)$ and articulates a vital scaling property.

LEMMA 5. *The function w has the following properties:*

(a) $w(t) \leq c_\psi^2 t$ *for* $0 \leq t \leq 1$,

(b) *For* $0 \leq t \leq q = 1/p$, $|m(t) - w(t)| \leq c_\psi^2 t^2$,

(c) $w(t) = pw(t/p)$.

Proof. By Lemma 3, we have $m(u) \leq c_\psi^2 u$ for $0 \leq u \leq 1$; so, for $0 \leq x \leq p^n q$ we have

$$f_n(x) = p^n m(x, p^n) \leq p^n c_\psi^2 x/p^n = c_\psi^2 x, \tag{14}$$

yielding (a). To show (b), we note that the case where $n = 0$ in inequality (7) states that, for $0 \leq x \leq q$,

$$m(x) = f_0(x) \leq m(x) \leq g_0(x),$$

and, by the definition of g_0, we obtain

$$|m(x) - w(x)| \leq |g_0(x) - f_0(x)| = c_\psi^2 x^2, \qquad 0 \leq x \leq q.$$

All that remains is to establish (c). By Lemma 4 and the definition of g_0, we have

$$\begin{aligned} w(x/p) &= \lim_{n\to\infty} p^n m\left(\frac{x}{p \cdot p^n}\right) \\ &= \lim_{n\to\infty} \frac{1}{p}\left\{p^{n+1} m\left(\frac{x}{p^{n+1}}\right)\right\} \\ &= w(x)/p, \end{aligned}$$

completing the proof of the lemma.

LEMMA 6 (first Laplace representation). *As $n \to \infty$, we have*

$$ES_n = n^2 \int_0^1 m(t) e^{-nt} dt + O(1/n). \tag{15}$$

Proof. The difference between ES_n and the integral of (15) is bounded by

$$\int_0^1 m(t)|n^2e^{-nt} - n(n-1)(1-t)^{n-2}|dt,$$

and, by Lemma 3, $m(t) = O(t)$ as $t \to 0$; so easy estimates give the lemma.

We can modify this last representation slightly to obtain one with the form of a standard Laplace integral. First, we note that

$$\begin{aligned} ES_n &= n^2 \int_0^q e^{-nt}m(t)dt + n^2 \int_q^1 e^{-nt}m(t)dt + O(1/n) \\ &= n^2 \int_0^q e^{-nt}w(t)dt + O(1/n), \end{aligned}$$

since

$$\int_0^q n^2e^{-nt}|m(t) - w(t)|dt \le \int_0^q n^2e^{-nt}c_\psi^2t^2dt = O(1/n)$$

and

$$n^2 \int_q^1 m(t)e^{-nt}dt \le n^2 \int_q^1 c_\psi^2te^{-nt}dt \le c_\psi n^2e^{-nq}.$$

Moreover, we have

$$\int_q^\infty n^2w(t)e^{-nt}dt \le n^2 \int_q^\infty c_\psi^2te^{-nt}dt = O(n^2e^{-qn});$$

so we have proved the following lemma.

LEMMA 7 (second Laplace representation). *It holds that*

$$ES_n = n^2 \int_0^\infty w(t)e^{-nt}dt + O(1/n).$$

All that remains is to show that the last integral is a periodic function of $\log_p n$. If we let $I(n)$ denote the value of the integral, divide the interval $[0, \infty)$ into subintervals $[p^k/n, p^{k+1}/n]$, where $-\infty < k < \infty$, and let $t = p^{k+u}/n$, then we obtain

$$\begin{aligned} I(n) &= n^2 \sum_{k=-\infty}^{\infty} \int_{p^k/n}^{p^{k+1}/n} w(t)e^{-nt}dt \\ &= n \sum_{k=-\infty}^{\infty} \int_0^1 w(p^{k+u}/n)p^{k+u} \exp(-p^{k+u}) \log p du. \end{aligned} \tag{16}$$

Using the key recursion relation of Lemma 5, part (c), we have $w(p^{k+u}/n) = p^kw(p^u/n)$; so

$$\begin{aligned} I(n) &= \log p \int_0^1 w(p^{u-\log_p n})/(p^{u-\log_p n}) \left\{ \sum_{k=-\infty}^{\infty} p^{2(k+u)} \exp(-p^{k+u}) \right\} du \\ &= \log p \int_0^1 \left\{ w(p^{u-\log_p n})/(p^{u-\log_p n}) \right\} l(u)du, \end{aligned}$$

where $l(u)$ is defined by

$$l(u) = \sum_{k=-\infty}^{\infty} p^{2(k+u)} \exp(-p^{k+u}).$$

Since the defining sum for l converges uniformly on compact subsets of $[0, \infty)$, we see that l is continuous. Furthermore, if we define a function φ by

$$\varphi(x) = \log p \int_0^1 \left\{ w(p^{-(x-u)})/(p^{-(x-u)}) \right\} l(u) du, \tag{17}$$

the continuity of l, the local boundedness of w, and the convolution form of (17) show that φ is also continuous. Since $\varphi(\log_p n) = I(n)$, we can write Lemma 7 as

$$ES_n = \varphi(\log_p n) + O(1/n), \tag{18}$$

and the proof of the theorem is completed once we establish the following lemma.

LEMMA 8. *The function φ is continuous, periodic of period* 1, *and bounded away from zero.*

Proof. We have already noted the continuity of φ, and periodicity is immediate from the recursion $w(pu) = pw(u)$ combined with the integral representation (17) of φ. The integral representation also gives $\varphi(x) \geq 0$ for all x.

By compactness, φ will be bounded away from zero unless there is an x_0 such that $\varphi(x_0) = 0$. However, for such an x_0, we would get from (17) that $w(p^{-(x_0-u)}) = 0$ for all $0 \leq u \leq 1$. By the recursion $w(pu) = pw(u)$, we could then conclude that both φ and ES_n were identically zero. This contradiction establishes that $\varphi(x)$ is bounded away from zero. By (18) and division by $\varphi(\log_p n)$, we find that

$$\frac{ES_n}{\varphi(\log_p n)} = 1 + O(1/n)$$

which is more than we require to complete the proof of the theorem.

4. Beyond expectations. One finds no difficulty in extending Theorem 1 beyond the convergence of expectations to almost sure convergence. The first step is to obtain an understanding of $VarS_n$, and this is easily approached though the use of martingales. For the martingale difference sequence defined by $d_i = E(S_n|\mathcal{F}_i) - E(S_n|\mathcal{F}_{i-1})$ where $\mathcal{F}_i = \sigma\{X_1, X_2, \ldots, X_n\}$, we have the representation

$$S_n - ES_n = \sum_{i=1}^{n} d_i,$$

and, by the orthogonality of the martingale differences, we have

$$VarS_n = E|S_n - ES_n|^2 = \sum_{i=1}^{n} Ed_i^2.$$

To help estimate Ed_i^2, we introduce another collection of random variables $\{\tilde{X}_i, 1 \leq i \leq n\}$ that are assumed to be independent, uniformly distributed, and independent of the random variables $\{X_i, 1 \leq i \leq n\}$. We then let $S_n^{(i)}$ denote the sum of squares of the edges in the spacefilling heuristic tour of $\{X_1, X_2, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n\}$ and note by $E(S_n^{(i)}|\mathcal{F}_i) = E(S_n|\mathcal{F}_{i-1})$ that we have

$$d_i = E(S_n - S_n^{(i)}|\mathcal{F}_i).$$

The basic observation is that the tours associated with S_n and $S_n^{(i)}$ differ by at most six edges. Furthermore, if we recall the oriented distance function $d(s,t)$ defined in §3 and specify integers j_1 and j_2 by the relations

$$d(t_{j_1}, \tilde{t_i}) = \min_{j:j\neq i} d(t_j, \tilde{t_i}) \quad \text{and} \quad d(\tilde{t_i}, t_{j_2}) = \min_{j:j\neq i} d(\tilde{t_i}, t_j),$$

then we obtain the spacefilling tour through $\{X_1, X_2, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n\}$ by connecting $\tilde{X}_i$ into the spacefilling tour through $\{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$ by edges to X_{j_1} and X_{j_2} and removing the edge that connects X_{j_1} and X_{j_2}. We are thus led to

$$\begin{aligned} S_n^{(i)} = S(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) &+ ||\psi(t_{j_1}) - \psi(\tilde{t_i})||^2 \\ + ||\psi(\tilde{t_i}) - \psi(t_{j_2})||^2 &- ||\psi(t_{j_1}) - \psi(t_{j_2})||^2. \end{aligned}$$

Similarly, to build the spacefilling tour through $\{X_1, X_2, \ldots, X_n\}$ from the tour through $\{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$, we find k_1 and k_2 such that

$$d(t_{k_1}, t_i) = \min_{k:k\neq i} d(t_k, t_i) \quad \text{and} \quad d(t_i, t_{k_2}) = \min_{k:k\neq i} d(t_i, t_k);$$

from which we obtain

$$\begin{aligned} S_n = S(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) &+ ||\psi(t_{k_1}) - \psi(t_i)||^2 \\ + ||\psi(t_i) - \psi(t_{k_2})||^2 &- ||\psi(t_{k_1}) - \psi(t_{k_2})||^2. \end{aligned}$$

The implied bound on $|S_n^{(i)} - S_n|$ is then

$$\begin{aligned} |S_n^{(i)} - S_n| \leq ||\psi(t_{j_1}) - \psi(\tilde{t_i})||^2 &+ ||\psi(\tilde{t_i}) - \psi(t_{j_2})||^2 \\ + ||\psi(t_{j_1}) - \psi(t_{j_2})||^2 &+ ||\psi(t_{k_1}) - \psi(t_i)||^2 \\ + ||\psi(t_i) - \psi(t_{k_1})||^2 &+ ||\psi(t_{k_1}) - \psi(t_{k_2})||^2, \end{aligned}$$

from which we find by Assumption 4 that $E|d_i|^2 \leq E|S_n - S_n^{(i)}|^2$ is bounded by

(19)
$$6c_\psi^2 E\left(\rho^2(t_{j_1}, \tilde{t_i}) + \rho^2(\tilde{t_i}, t_{j_2}) + \rho^2(t_{j_1}, t_{j_2}) + \rho^2(t_{k_1}, t_i) + \rho^2(t_i, t_{k_2}) + \rho^2(t_{k_1}, t_{k_2})\right).$$

The computation of the expectations in (19) are now routine, given the known distribution of the gaps between n (or $n-1$) points chosen on the unit circle. All of these expectations are $O(1/n^2)$, and hence we have Lemma 9.

LEMMA 9. *For $n \to \infty$, the variance of the sum of squares of edges S_n satisfies*

$$Var S_n = O(1/n).$$

Now we are ready to prove the almost sure convergence of S_n.

THEOREM 2. *We have*

$$\lim_{n\to\infty} \frac{S_n}{\varphi(\log_p n)} = 1 \quad a.s.$$

Proof. The remaining steps follow a familiar pattern and will just be sketched. First, we consider a subsequence by letting $n_i = \lceil i \log^2 i \rceil$. By Chebyshev's inequality, for any $\epsilon > 0$, we have

$$\sum_{i=1}^{\infty} P(|S_{n_i} - ES_{n_i}| > \epsilon) \leq \epsilon^{-2} \sum_{i=2}^{\infty} \mathrm{Var} S_{n_i} \leq \epsilon^{-2} \sum_{i=2}^{\infty} i^{-1} \log^{-2} i < \infty;$$

so, by the easy part of the Borel–Cantelli lemma, for $n \to \infty$, we have $(S_{n_i} - ES_{n_i}) \to 0$ with probability 1. To stretch the convergence to the full sequence, we look at the largest difference V_i of S_n and S_{n_i} as n varies through the intervals $[n_i, n_{i+1})$,

$$V_i = \max_{n_i \leq n \leq n_{i+1}} |S_n - S_{n_i}|.$$

For each n such that $n_{i-1} \leq n \leq n_i$, the number of terms that occur in S_n and not in S_{n_i} or vice versa is at most $O(n_i - n_{i-1}) = O(\log^2 i)$. If A denotes the set of all such terms, then $|A| = O(\log^4 i)$; and, if L_i is the largest of these terms, then by Assumption 4 and the classical proof of the fact that the largest gap between n points chosen at random from the circle is sharply concentrated around $n^{-1} \log n$, we can show that $EL_i^2 = O(n_i^{-2} \log^2 n_i)$. Since $V_i \leq |A| L_i$, we see $EV_i^2 = O(i^{-2} \log^{10} i)$; so, again by the Borel–Cantelli lemma, we see that, for $n \to \infty$, we have $V_i \to 0$ with probability 1, which, in turn, completes the proof of the theorem.

5. Concluding remarks. We have established a striking property of the sum of squares of the heuristic tour provided by the spacefilling curve method. Of the questions that remain open, the most natural are perhaps those that seek a more detailed understanding of φ, the periodic function that figures in Theorem 1. The (centered) supremum norm of φ determines the strength of the oscillation of ES_n, and almost nothing is known about this norm. Simulations offer little insight because of the slow growth of $\log_p n$ and the difficulty of estimating ES_n. Still, calculations given in Platzman and Bartholdi [8] would suggest by analogy that the oscillation due to φ is not large, perhaps only a few percent. This reinforces the difficulty of obtaining detailed information about φ from simulations.

Another natural question addresses the possibility of obtaining sharp bounds on the tail probabilities $P(S_n \geq t)$. In view of the remarkable work of Rhee and Talagrand [9] on the Gaussian tail bound for the optimal length of the TSP tour length, we are tempted to suggest that $P(S_n \geq t) \leq A \exp(-Bt^2)$ for all some A and B and all $t \geq 0$. Some of the structure that Rhee and Talagrand require (like the martingale used in §4) is available in for S_n, but a detailed understanding of $P(S_n \geq t)$ seems to require additional insights, since, in particular, S_n may have tails that are much lighter than those of the Gaussian distribution.

Perhaps the most compelling problems suggested by this work concern the behavior of the *optimal tours* rather than the heuristic cousins. First, if we let S_n^{TSP} denote the sum of the squares of the edges in the (almost surely unique) shortest tour through a random sample of n points chosen from the unit square, do we have

$$\lim_{n\to\infty} ES_n^{\mathrm{TSP}} = C$$

for some constant C?

Finally, there at least one compelling problem concerning the worst-case behavior of the sum of squared edges in an minimal length path. If S is a finite subset of $[0,1]^2$ and $\mathrm{TSP}(S)$ denotes a path of minimal length through the points of S, we suspect that

$$M_n = \max_{S:|S|=n} \sum_{e \in \mathrm{TSP}(S)} ||e||^2$$

is asymptotic to $c \log n$ as $n \to \infty$. By the results cited in the first section, we know $M_n = O(\log n)$ and $M_n = \Omega(\log n)$, but the present methods offer no serious progress toward a full asymptotic result. The periodicity that has been found for ES_n shows that subtleties can emerge, though we need not expect them at every turn.

REFERENCES

[1] D. ALDOUS AND J. M. STEELE, *Asymptotics for Euclidean minimal spanning trees on random points,* Probab. Theory Rel. Fields, 92 (1992), pp. 247–258.

[2] J. BEARDWOOD, J. H. HALTON, AND J. M. HAMMERSLEY, *The shortest path through many points*, Proc. Cambridge Philos. Soc., 55 (1959), pp. 299–327.

[3] M. BERN AND D. EPSTEIN, *Worst-case bounds for subadditive geometric graphs*, Tech. Report, Xerox PARC, Palo Alto, CA, 1992.

[4] J. GAO AND J. M. STEELE, *General spacefilling curve heuristics and limit theory for the traveling salesman problem*, J. Complexity, 10 (1994), to appear.

[5] D. HILBERT, *Über die stetige Abbildung einer Linie auf ein Flächenstück*, Math. Ann., 38 (1891), pp. 459–460.

[6] S. C. MILNE, , *Peano curves and smoothness of functions*, Adv. Math., 35 (1980), pp. 129–157.

[7] G. PEANO, *Sur Une Courbe Qui Remplit Toute Une Aire Plane*, Math. Ann., 36 (1890), pp. 157–160.

[8] L. K. PLATZMAN AND J. J. BARTHOLDI, *Spacefilling curves and the planar traveling salesman problem*, J. Assoc. Comput. Mach., 36 (1989), pp. 719–737.

[9] W. T. RHEE AND M. TALAGRAND, *A sharp deviation inequality for the stochastic traveling salesman problem*, Ann. Probab., 17 (1989), pp. 1–8.

[10] T. L. SNYDER AND J. M. STEELE, *A priori inequalities for the Euclidean traveling salesman problem*, in Proc. 8th ACM Sympos. on Computational Geometry, Berlin, 1992, pp. 344–349.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 325–330, May 1994

EULERIAN SELF-DUAL CODES*

L. BABAI[†], H. ORAL[‡], AND K.T. PHELPS[§]

Abstract. The authors present a construction of binary self-dual codes from Eulerian graphs and establish that the code will be indecomposable if and only if the vertices of degree 2 are not a cutset of the graph. The construction is used to establish that every finite group is isomorphic to the automorphism group of some self-dual code. It is further shown that deciding isomorphism of self-dual codes is at least as difficult as graph isomorphism.

Key words. Eulerian graph, self-dual code, automorphisms

AMS subject classifications. 94B05, 05C99

1. Introduction.

There are many constructions of codes from the incidence matrices associated with graphs, designs, and so forth. While in most cases the codes are not especially *good*, the structure of the combinatorial object allows us to establish results about the corresponding code. For example, the vertex-edge incidence matrix of a connected graph X generates a code of length $|E(X)|$ and dimension $|X| - 1$. Nonisomorphic graphs may generate isomorphic codes. However, as a consequence of a famous result of Whitney, the codes of 3-connected graphs are isomorphic if and only if the graphs are. One consequence of this construction of codes is that every finite group is the automorphism group of some linear code. Another consequence is that deciding on the isomorphism of linear codes is at least as difficult as deciding on the isomorphism of graphs. In this paper, we intend to establish similar results for self-dual codes, but first we present a construction of self-dual codes from Eulerian graphs.

Before proceeding with the construction, we introduce some definitions. First, a binary code of length n and dimension k is just a linear subspace of V^n, the vector space of dimension n over $GF(2)$. A code is said to be self-dual if it is equal to its dual. Of course, this means that the length n is even and that the dimension is $n/2$. A self-dual code is said to be *indecomposable* if it is not isomorphic to the direct sum of two shorter self-dual codes.

A connected graph X is Eulerian if all vertices have even degree. If the graph is not connected but each component is Eulerian, our construction still applies. However, the self-dual code constructed in this manner is clearly decomposable.

There is at least one other construction of self-dual codes from graphs: The face-vertex incidence matrix of a cubic planar bipartite graph generates a self-dual code [2]. However, the structure of these codes in terms of automorphism groups and others is limited.

2. Eulerian self-dual codes.

Let X be an Eulerian graph on n vertices $\{x_1, \ldots, x_n\}$ with no isolated vertices. For each edge $e = (x_i, x_j)$ of X, we intro-

* Received by the editors July 24, 1991; accepted for publication (in revised form) March 23, 1993.

[†] Department of Computer Science, University of Chicago, Chicago, Illinois 60637 (laci@cs.uchicago.edu).

[‡] Mathematics Department, Bosphorus University, Bebek, Istanbul, Turkey (oral@trboun.bitnet). This research was supported by North Atlantic Treaty Organization CRW grant 900976 and TUBITAK project TBAG-1082.

[§] Discrete and Statistical Sciences, Auburn University, 120 Mathematics Annex, Auburn, Alabama 36849-5307 (ktphelps@ducvax.auburn.edu). This research was supported by North Atlantic Treaty Organization CRW grant 900976.

duce two half edges or flags, denoted by $[e, x_i]$ and $[e, x_j]$. We define the edge-flag incidence matrix of X as follows.

DEFINITION. *Let $A = [a_{ij}]$ be the $|E(X)| \times 2|E(X)|$ binary incidence matrix whose rows are indexed by the edges of X and columns are indexed by the flags (half edges). An edge e' is incident with a flag $[e, x_i]$ if $e' \neq e$, and e' is incident with x_i.*

We henceforth identify the rows of A by the edges of X. Note that the *weight* of any row of A is just the number of 1's in it. We now can prove the following result.

THEOREM 2.1. *Let X be an Eulerian graph on n vertices and let A be the edge-flag incidence matrix of X. Then A is a generator matrix of a binary self-dual code.*

Proof. We must prove that the rows of A are orthogonal to each other and that they are linearly independent. Let A_e be row of A corresponding to the edge $e = (x_i, x_j)$ of X; then the weight of A_e is equal to $d(x_i) + d(x_j) - 2$, which is even. So any row of A is orthogonal to itself. Let A_u and A_v be two distinct rows of A corresponding to the edges (x_i, x_j) and (x_k, x_l). If x_i, x_j, x_k, x_l are all distinct, then A_u and A_v have no 1's in common, and so they are orthogonal. If they share a vertex (say) $x_i = x_k$, then A_u and A_v have $d(x_i) - 2$ 1's in common. Again, since $d(x_i) - 2$ is even, this implies the orthogonality of A_u and A_v.

Now we prove that the rows of A are linearly independent. Assume that we can find a subset B of edges of X such that $\sum_{e \in B} A_e = 0$. Let x_i be a vertex of X with degree d_i; then x_i corresponds to d_i flags (half edges) and thus d_i columns of A. The nonzero entries in these columns of A occur in the rows that correspond to the edges of X incident with the vertex x_i. This $d_i \times d_i$ submatrix of the matrix A is just $J_{d_i} - I_{d_i}$, where J denotes all-1's matrix, I denotes the identity matrix, and the subscripts indicate the dimension. Since this submatrix has full rank, there can be no edges in B incident with x_i, and thus B must be empty. We deduce that any nonempty subset of the set of rows of A is linearly independent, and thus A has full rank. Therefore A is a generator matrix of a self-dual code of length $2|E(X)|$. □

Let X be an Eulerian graph and let C the corresponding self-dual code. Each codeword of C corresponds to an edge-induced subgraph of X. Let u be an element of C and let U be the corresponding subgraph of X. Let x be a vertex in U. By $d_U(x)$, we denote the degree of x in the subgraph U. As we noted in the proof of Theorem 2.1, each vertex $x_i \in X$ corresponds to $d(x_i)$ columns of A and thus $d(x_i)$ coordinate places of any codeword u. Let $u(x)$ denote the coordinate places corresponding to the vertex $x \in X$. Then we have that the number of nonzero entries in $u(x)$ is $d_U(x)$ or $d(x) - d_U(x)$, depending on whether $d_U(x)$ is even or odd, respectively. To see this, observe that the sum of k rows of $J_{d_i} - I_{d_i}$ has weight k if k is even and if weight $d_i - k$ if k is odd. Obviously, the weight of u is just the sum of the weights of the $u(x), x \in V(U)$, and these weights are always positive. Using the above explanation and Theorem 2.1, we can prove the following theorem.

THEOREM 2.2. *Every finite group is isomorphic to the automorphism group of some binary self dual code.*

Proof. Let X be a 4-regular graph of girth 7 on n vertices and let A be the edge-flag incidence matrix of X. By Theorem 2.1, we know that the row space of A is a self-dual code C of length $4n$. It is relatively easy to show that any finite group occurs as the automorphism group of an 4-regular graph of girth 7 (e.g., see [5]). If we prove that the automorphism group of C is isomorphic to the automorphism group of X, we are done. To prove this fact, first we prove that C has no element of weight 6 other than the rows of A.

Each vertex of X has degree 4, so the weight of each row of A is 6. Since the rows of A are linearly independent, then there is a one-to-one correspondence between edge-induced subgraphs of X and the elements of $\mathcal{C}$, namely,

$$U \subset X \longleftrightarrow \sum_{e \in E(U)} A_e$$

(where A_e is the row of A corresponding to the edge $e \in U$ and where the summation is modulo 2).

Let u be an element of $\mathcal{C}$ and let U be the corresponding subgraph of X. Let x be a vertex in U. By $d_U(x)$, we denote the degree of x in the subgraph U. We assume that U is not an edge, so it has at least two edges and hence at least three vertices. If the number of vertices of U is less than 7, then U is a forest, and hence it has at least two vertices of degree 1. Since the vertices of degree 1 contribute 3 to the weight of u, the weight of u is strictly greater than 6. If U has at least 7 vertices, then the weight of u is at least 7, since the contribution of each vertex is positive. Thus the only elements of weight 6 are the codewords corresponding to the edges of X.

Since the only codewords of weight 6 are the rows of A, any permutation of the columns of A is an automorphism of $\mathcal{C}$ if and only if it induces a permutation of the rows of A (i.e., the edges of X) and thus is an automorphism of X. This proves our theorem. □

We mentioned that, if the graph is not connected but each component is Eulerian, then our construction would give a decomposable self-dual code. We next characterize graphs that give indecomposable codes. In particular, we prove that the code we obtain with this construction is indecomposable if and only if the graph does not contain a set vertices of degree 2 that is a cutset. Obviously, if the graph is not connected, then the empty set is such a cutset, and thus it is sufficient to prove this for Eulerian graphs.

To prove this result, we first note that codewords corresponding to the edges of the graph are minimal. Let $\text{Sup}(u) = \{i | u_i = 1\}$ denote the support of a codeword u; then $u \in \mathcal{C}$ is minimal if there is no nonzero codeword in $\mathcal{C}$ whose support is a proper subset of $\text{Sup}(u)$. Since each codeword corresponds to a subgraph and each vertex of positive degree in this subgraph contributes to the support, if $\text{Sup}(w) \subset \text{Sup}(u)$ for $u, w \in \mathcal{C}$, then $V(W) \subset V(U)$, where W, U are the subgraphs corresponding to w, u and $V(W), V(U)$ are the vertex sets of these edge-induced subgraphs. In particular, this means that, assuming there are no multiple edges, the words corresponding to edges are always minimal.

THEOREM 2.3. *Let X be an Eulerian graph and let $\mathcal{C}$ the self-dual code constructed from it; then $\mathcal{C}$ is indecomposable if and only if no subset of vertices of degree 2 in X is a cutset.*

Proof. Let X be an Eulerian graph and let A the corresponding edge-flag incidence matrix as constructed above; let $\mathcal{C}$ be the self-dual code having A as its generator matrix. Since the rows of A are minimal codewords, the code $\mathcal{C}$ is decomposable if and only if the matrix A is, that is, if and only if A is equivalent to

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}.$$

Note that, since the rows of A are independent and mutually orthogonal, both A_1 and A_2 generate self-dual codes.

For any vertex $x \in X$ having degree $\text{d}(x) > 2$, the corresponding $d(x)$ columns in A must all be contained in the same part. On the other hand, the two columns of

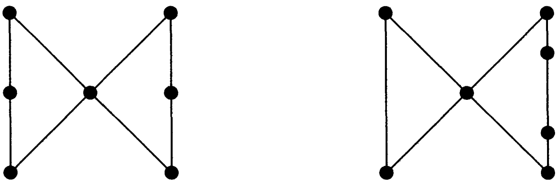


FIG. 1. *Graphs that produce isomorphic codes.*

A corresponding to a vertex of degree 2 in X can be in different parts, as they have only one nonzero entry. If a set of vertices of degree 2 is a (minimal) cutset in X, then this induces a decomposition of the matrix A. Conversely, any decomposition of A can only split sets of columns corresponding to vertices of degree 2, and thus these vertices must form a cutset in X. □

3. Isomorphism of Eulerian self-dual codes. While isomorphic Eulerian graphs produce isomorphic self-dual codes with the above construction, the converse is not true in general. However, we show that it is true for a class of graphs that are *isomorphism complete*, that is, equivalent to the graph isomorphism problem.

In Fig. 1, we give two Eulerian graphs that give isomorphic self-dual codes. The codes are decomposable. An indecomposable self-dual code of length 20 can be constructed from both K_5, the complete graph on five vertices, and the graph on six vertices in Fig. 2. Consider the code constructed from K_5; it has exactly five disjoint codewords of weight 4 corresponding to the five subgraphs K_4. Figure 2 has four words of weight 4 that correspond to edges and one subgraph K_4 that gives the fifth word of weight 4. Since there is only one such indecomposable code of length 20, which is called M_{20} in [4], we conclude that they are isomorphic. We give the generator matrix constructed from K_5 in Fig. 3.

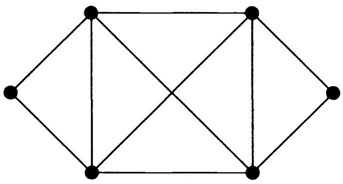


FIG. 2. *Graph on ten edges.*

We now consider the isomorphism problem for self-dual codes, which is as follows: Given two self-dual codes of length n, is there a permutation matrix that maps one subspace to the other? If we represent the code by an explicit list of all codewords, then the isomorphism question is equivalent to graph isomorphism. However, linear codes are usually represented by a generator matrix (or basis), which is logarithmic in the size of the code. Next, we intend to establish that the isomorphism problem for graphs reduces to the isomorphism problem for self-dual codes.

There are various classes of graphs that are known to be isomorphism complete: regular graphs, Eulerian graphs, regular bipartite graphs, and so forth [1]. We claim, in fact, that the following result can be easily proved using the same techniques.

LEMMA 3.1. *Regular Eulerian bipartite graphs of degree r containing no subgraph isomorphic to the complete bipartite graph $K_{r-1,r-1}$ are isomorphism complete.*

We now show that two graphs from this class are isomorphic if and only if the corresponding self-dual codes are; thus, deciding graph isomorphism reduces to deciding isomorphism of self-dual codes.

$$\left|\begin{array}{cccccccccccccccccccc}
0&1&1&1&0&1&1&1&&&&&&&&&&&&\\
1&0&1&1&&&&&0&1&1&1&&&&&&&&\\
1&1&0&1&&&&&&&&&0&1&1&1&&&&\\
1&1&1&0&&&&&&&&&&&&&0&1&1&1\\
&&&&1&0&1&1&1&0&1&1&&&&&&&&\\
&&&&1&1&0&1&&&&&1&0&1&1&&&&\\
&&&&1&1&1&0&&&&&&&&&1&0&1&1\\
&&&&&&&&1&1&0&1&1&1&0&1&&&&\\
&&&&&&&&1&1&1&0&&&&&1&1&0&1\\
&&&&&&&&&&&&1&1&1&0&1&1&1&0
\end{array}\right|$$

FIG. 3. *Self-dual code from K_5.*

THEOREM 3.2. *Let X and X^* be two regular Eulerian bipartite graphs of degree r having no sub-$K_{r-1,r-1}$ and let C and C^* be the corresponding self-dual codes constructed from them as above. Then these graphs are isomorphic if and only if the self-dual codes are.*

Proof. Let X be a regular bipartite graph of degree r, where r is even, and assume that it does not contain a sub-$K_{r-1,r-1}$. Let C be the corresponding self-dual code. We intend to show that the only codewords in C of weight $2r-2$ are those corresponding to edges in X. First, note that $2r-2 \equiv 2 \pmod 4$, since r is even. If U is an edge-induced subgraph of X whose vertex degrees are all even, then the corresponding codeword u has weight $wt(u) \equiv 0 \pmod 4$. To see this, recall that $wt(u)$ is just the sum of the degrees in this case; however, a bipartite graph with all vertex degrees even has an even number of edges, and hence the sum of the degrees is divisible by 4.

Now consider subgraphs U having $2k$ odd vertices, $k \geq 1$. We assume that the vertices of odd degree are $\{x_1, x_2, \ldots, x_{2k}\}$. If u is the corresponding codeword, then

$$wt(u) = \sum_{i=1}^{2k} r - d_U(x_i) + \sum_{i=2k+1}^{n} d_U(x_i).$$

We wish to minimize this expression over all subgraphs $U \subset X$ having at least two edges and least two vertices of odd degree. Assume that we have $2k = k_1 + k_2$ vertices of odd degree with k_1 in one part and k_2 vertices in the other. The subgraph that minimizes $wt(u)$ has as many edges as possible between vertices of odd degree. In this case, we always obtain

$$wt(u) \geq 2kr - 2k_1k_2,$$

which is minimal when $k_1 = k_2 = k$. On the interval $(1, r-1)$, the expression $2kr-2k^2$ has its minimum values at the endpoints, and thus

$$wt(u) > 2r - 2$$

for these values of k as well as for $k \geq r$. For $k = r - 1$, we have assumed that no sub-$K_{r-1,r-1}$ exists, so some of the vertices of odd degree must be adjacent to vertices of even degree, and thus again $wt(u) > 2r - 2$. For the last case where $k = 1$, the minimum occurs when U is a path of length 2 and again $wt(u) > 2r - 2$. □

4. Final comments. The results of the previous sections lead to several interesting questions. First, is there a reasonable characterization of classes of Eulerian graphs whose corresponding self-dual codes are unique to within isomorphism? For graphical codes, we have that connectivity is the key issue. In particular, graphical codes is isomorphic if and only if the corresponding graphs are 2-isomorphic. In the case of 3-connected graphs, 2-isomorphism reduces to just isomorphism.

Second, is deciding isomorphism of self-dual codes isomorphism complete? Of course, this remains unsolved even for binary linear codes in general. We remark that isomorphism of binary linear codes reduces to isomorphism of permutation groups (given a list of generators). Isomorphism of permutation groups is known to be in NP as well as the complexity class CoAM. This latter fact implies that the problem is not NP-complete unless the polynomial time hierarchy of complexity classes collapses (it is conjectured not to).

Finally, let us comment on the minimum distance. It was shown [3] that a self-dual code with minimum distance 4 cannot have trivial automorphism group. So, to construct a self-dual code with any automorphism group, we should have distance at least 6, as we had in this paper.

REFERENCES

[1] K. S. Booth and C. J. Colbourn, *Problems polynomial equivalent to graph isomorphism*, Report cs-77-04, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, June 1979.

[2] H. Oral, *Constructing self-dual codes using graphs*, J. Combin. Theory Ser. B, 52 (1991), pp. 250–258.

[3] ———, *Self-dual codes of distance four are not rigid*, Discrete Math., 104 (1992), pp. 263–271.

[4] V. Pless, *A classification of self-dual codes over $GF(2)$*, Discrete Math., 3 (1972), pp. 209–246.

[5] G. Sabidussi, *Graphs with given group and given graph-theoretical properties*, Canad. J. Math., 9 (1957), pp. 515–527.

SIAM J. DISCRETE MATH.
Vol. 7, No. 2, pp. 331–336, May 1994

020

CLIQUE GRAPHS OF CHORDAL AND PATH GRAPHS*

JAYME L. SZWARCFITER† AND CLAUDSON F. BORNSTEIN‡

Abstract. Clique graphs of chordal and path graphs are characterized. A special class of graphs named **expanded trees** is discussed. It consists of a subclass of disk-Helly graphs. It is shown that the clique graph of every chordal (hence path) graph is an expanded tree. In addition, every expanded tree is the clique graph of some path (hence chordal) graph. Different characterizations of expanded trees are described, leading to a polynomial time algorithm for recognizing them.

Key words. algorithms, chordal graphs, clique graphs, path graphs

AMS subject classifications. 05C12, 05C75, 05C85

1. Introduction. We examine clique graphs of chordal graphs. Bandelt and Prisner [2] proved that they are disk-Helly. Chen and Lih [4] and, independently, Bandelt and Prisner [2] showed that the second iterated clique graph of a chordal graph is again chordal. Here it is shown that clique graphs of chordal graphs correspond to a class, named expanded trees, in the sense that the clique graph of a chordal graph is always an expanded tree and that every expanded tree is the clique graph of some chordal graph. In addition, the class of clique graphs of (undirected) path graphs is no more restricted than that of chordal graphs. Every expanded tree is also the clique graph of some path graph. Expanded trees are characterized, and a polynomial time recognition algorithm is described.

Expanded trees are closely related to dismantlable graphs. The latter were examined by Bandelt and Prisner [2], Prisner [9], and Nowakovski and Winkler [8]. Disk-Helly graphs are a subclass of dismantlable graphs and can be recognized in polynomial time, according to an algorithm by Bandelt and Pesch [1]. See also Nowakovski and Rival [7] and Quilliot [10].

G denotes a simple undirected graph, $V(G)$ and $E(G)$ are its vertex and edge sets, respectively, $n = |V(G)|$ and $m = |E(G)|$. $N(v)$ is the set of vertices adjacent to $v \in V(G)$, while $N[v] = N(v) \cup \{v\}$. The vertex $v \in V(G)$ is *dominated* by $w \in V(G)$ *in* G when v, w are distinct and $N[v] \subset N[w]$. A *clique* is a subset of vertices inducing a complete subgraph in G. Let F be a family of subsets of some set. The *intersection graph* of F is a graph whose vertices are associated with the subsets of F, two vertices being adjacent if the corresponding pair of subsets intersect. The *clique graph* $K(G)$ of G is the intersection graph of the maximal cliques of G. A *chordal graph* G is the intersection graph of subtrees of a tree T. The subtree of T corresponding to a vertex $v \in V(G)$ is called *representative subtree* of v and denoted by $T(v)$. The tree T, together with the representative subtrees, forms a *tree representation* of G. A *minimal representation* is a tree representation such that $|V(T)|$ is the least possible. Gavril [5] and Buneman [3] showed that a minimal representation is precisely one in which each vertex of T corresponds to a maximal clique of G. In addition, for each

* Received by the editors December 9, 1991; accepted for publication (in revised form) April 12, 1993. The authors are on leave from Universidade Federal do Rio de Janeiro, Instituto de Matemática and Núcleo de Computação Electrônica, Caixa Postal 2324, 20001 Rio de Janeiro, RJ, Brasil.

† Current address, Universitè de Paris-Sud, Laboratoire de Recherche en Informatique, Bât. 490, 91405 Orsay Cedex, France. This author was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Brazil grant 200.076-83.

‡ Current address, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213. This author was supported by Comissão de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES) Brazil grant 9005009.

$v \in V(G)$, the subtree $T(v)$ is formed exactly by the vertices of T corresponding to those maximal cliques of G that contain v. A *path graph* is the intersection graph of paths of a tree. Monma and Wei [6] characterized path graphs and variations of this class in terms of their minimal representations.

Let F be a family of subsets S_i of some set. F satisfies the *Helly property* when every subfamily F' of F in which $S_i \bigcap S_j \neq \emptyset$, for all pairs $S_i, S_j \in F'$, is such that $\bigcap_{S_i \in F'} S_i \neq \emptyset$. Finally, G is a *dismantlable graph* if there exists a sequence $v_1, \ldots, v_n$ of its vertices such that, for $i < n$, v_i is dominated in $G - \{v_1, \ldots, v_{i-1}\}$. If, additionally, the maximal cliques of G satisfy the Helly property, then G is *disk-Helly*.

2. Expanded trees. G is an *expanded tree* when it admits a spanning tree $T(G)$, such that, for each edge $(v,w) \in E(G)$, the vertices of the $v - w$ path in T form a clique in G. In this case, $T(G)$ is a *canonical tree* of G.

THEOREM 2.1. *The following are equivalent.*

(i) *G is the clique graph of some connected chordal graph H.*

(ii) *G admits a spanning tree T, such that, for each $v \in V(G)$, $N_G[v] \cap T$ is a (connected) subtree of T.*

(iii) *G is an expanded tree.*

Proof. (i)⇒(ii) Let H be a chordal graph, T a minimal representation of it, and $G = K(H)$. Let G' be the graph obtained from T by adding exactly the edges that transform each representative subtree $T(w)$ into a $|T(w)|$-clique. $V(G) \simeq V(G')$. Let M_1, M_2 be two maximal cliques of H and let v_1, v_2 and v'_1, v'_2 be their corresponding vertices in G and G', respectively. Suppose that $(v_1, v_2) \in E(G)$. Then there exists $w \in V(H)$ such that $w \in M_1 \cap M_2$. In consequence, the subtree $T(w)$ of the minimal representation of H contains $v'_1, v'_2 \in V(G')$. Hence $(v'_1, v'_2) \in E(G')$. Conversely, if $(v'_1, v'_2) \in E(G')$, then there is a representative subtree $T(w)$ containing v'_1, v'_2. That is, $w \in M_1 \cap M_2$, and, consequently, $(v_1, v_2) \in E(G)$. Hence $G \simeq G'$, and T is a spanning tree of G. For any $v \in V(G)$, each $u \in N_G[v]$ corresponds to a subtree of T containing v. Hence $N_G[v] \cap T$ is connected.

(ii)⇒(iii) It suffices to show that T is a canonical tree of G. Let $(v,w) \in E(G)$ and let $v = v_0, \ldots, v_r = w$ be the $v - w$ path in T. Suppose by induction that $\{v_0, \ldots, v_{r-1}\}$ is a clique of G, $r > 1$. Since v and w are adjacent and $N_G[v] \cap T$ is connected, it follows that $v_0, \ldots, v_{r-1}$ are all adjacent to w. Consequently, $\{v_0, \ldots, v_r\}$ is a clique of G; that is, G is an expanded tree.

(iii)⇒(i) Given an expanded tree G, we construct a connected chordal graph H such that $G = K(H)$. Let M be the set of maximal cliques of G. Define H as the intersection graph of the elements of $M \cup V(G)$. Denote by T a canonical tree of G. We show that each maximal clique $C \in M$ of G induces a subtree in T, which implies that H is the intersection graph of subtrees of a tree. Let v, z be two vertices of C and let P be the $v - z$ path in T. For each vertex w of C, the paths $w - v$ and $w - z$ cover P. Since (w,v) and (w,z) are edges of the expanded tree G, it follows that w is adjacent to every vertex of P. By maximality of C, each vertex of P belongs to C. Hence $C \cap T$ is a subtree of T, and, consequently, H is a chordal graph. It remains to show that $G = K(H)$. Clearly, each vertex v of T corresponds to a maximal clique of H, namely, that formed by the maximal cliques of G that contain v and by v itself. That is, $V(G) \simeq V(K(H))$. Let C_1, C_2 be two maximal cliques of H and let v_1, v_2 be their corresponding vertices in G. If $C_1 \cap C_2 \neq \emptyset$, then there is a maximal clique C of G such that $C \in V(H)$ is contained in $C_1 \cap C_2$. In other words, v_1, v_2 are vertices of G belonging to the same clique C. That is, $(v_1, v_2) \in E(G)$. Conversely, if $C_1 \cap C_2 = \emptyset$, then there is no maximal clique $C \in V(H)$ of G containing both $v_1, v_2 \in V(G)$. That is, $(v_1, v_2) \notin E(G)$, and, consequently, $G = K(H)$. □

COROLLARY 2.2. *G is an expanded tree if and only if it is the clique graph of some connected path graph.*

Proof. Let G be an expanded tree. Construct a path graph H such that $G = K(H)$. By Theorem 1, there exists a chordal graph F such that $G = K(F)$. Let T be a minimal representation of F. Substitute each representative subtree (of a vertex of F) with the collection of all paths between its leaves. The intersection graph of all those paths is a path graph H with minimal tree representation T and such that $K(H) = K(F)$. The converse is clear. □

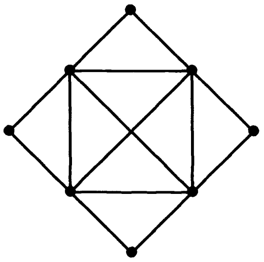


FIG. 1

Note the following relation between expanded trees and disk-Helly graphs. The clique graph of every chordal graph is disk-Helly [2]. Hence expanded trees are also disk-Helly, by Theorem 1. The inclusion is proper, as the graph of Fig. 1 is disk-Helly and not an expanded tree.

3. Recognition of expanded trees. A sequence S_k of vertices $v_1, \ldots, v_k$, $k \leq n$ of a graph G is *canonical* when, for each $1 \leq i \leq k$, either $i = n$ or v_i is dominated by some vertex v_j, $i < j$ in the graph $G(S_i)$, defined as

$$\begin{gathered} V(G(S_i)) = V(G), \\ E(G(S_i)) = E(G) - \{(x,y) \in E(G) \\ \text{s.t. } x \in \{v_1, \ldots, v_{i-1}\},\ y \in \{v_i, \ldots, v_n\}, \text{ and } |N_G(x) \cap \{v_i, \ldots, v_n\}| = 1\}. \end{gathered}$$

Call the vertex v_i *canonical* in $G(S_i)$. The value k is the *length* of S_k. If $k = n$, then S_k is *complete*. S_k is *maximal* when it is complete, or otherwise $G(S_{k+1})$ has no canonical vertex. Clearly, if S_k is canonical, then so is $S_i, i < k$. Define $G(S_0) = G$.

Expanded trees can also be characterized as follows.

THEOREM 3.1. *G is an expanded tree if and only if it admits a complete canonical sequence.*

Proof. ($\Rightarrow$) Let T be a canonical tree of G. Let S_n be a sequence $v_1, \ldots, v_n$ of the vertices of G, such that v_i is a leaf of T_i, $1 \leq i \leq n$, where $T_1 = T$ and, for $i > 1$, $T_i = T_{i-1} - v_{i-1}$. We show by induction that S_i is canonical. Assume it is true for all subsequences of length $< i$. If $i = n$, there is nothing to prove. Otherwise, let v_j, $j > i$ be the vertex adjacent to v_i in T_i. We claim that v_j dominates v_i in $G(S_i)$. This is clear for $i = 1$. When $i > 1$, suppose that the claim is false. Then there is a vertex v_p, such that $(v_p, v_i) \in E(G(S_i))$ and $(v_p, v_j) \notin E(G(S_i))$. T is canonical. Hence the following two conditions must hold: $p < i$ and $(v_p, v_q) \notin E(G)$ for all $q > i$; otherwise, (v_p, v_j) would belong to $E(G(S_i))$, a contradiction.

In this case, $|N_G(v_p) \cap \{v_i, \dots, v_n\}| = 1$, and $(v_p, v_i) \notin E(G(S_i))$, again a contradiction. Therefore v_j dominates v_i in $G(S_i)$; that is, S_i is canonical.

($\Leftarrow$) Let $v_1, \dots, v_n$ be a canonical sequence of G. Each v_i, $i < n$ has a dominator v_j in $G(S_i)$, $i < j$; write $v_j = \text{dom}(v_i)$. Let T be a graph defined as

$$V(T) = V(G),$$
$$E(T) = \{(v_i, \text{dom}(v_i)) \text{ s.t. } 1 \leq i \leq n\} \subset E(G).$$

When $n > 1$, every vertex of G is incident to some edge of T. Since $|E(T)| = n-1$, T is a spanning tree of G. We show that T is canonical. Suppose the contrary. Then G contains an edge (v_a, v_b) such that the $v_a - v_b$ path P in T is not a clique of G. We can choose (v_a, v_b) such that $(v_i, v_b) \in E(G)$ for all vertices v_i of P, except the neighbor v_c of v_a in P, which satisfies $(v_c, v_b) \notin E(G)$. Let v_d, $d \neq a, b$ be the second adjacent vertex to v_c in P. Let $\vec{T}$ be the in-tree obtained by directing each edge $(v_i, \text{dom}(v_i))$ of T from v_i to $\text{dom}(v_i)$.

Case 1. $a < b$.

The edge (v_a, v_c) must be oriented in $\vec{T}$ from v_a to v_c. Otherwise, $\vec{T}$ contains the directed path $v_b - v_a$, implying $b < a$, a contradiction. Then $a < c$, and $(v_a, v_c) \in E(\vec{T})$ implies that v_c dominates v_a in $G(S_a)$. Since $a < b$, $(v_a, v_b) \in E(G(S_a))$. Hence $(v_b, v_c) \in E(G)$, a contradiction.

Case 2. $a > b$.

Suppose that (v_a, v_c) is directed from v_c to v_a in $\vec{T}$. Then P is directed from v_b to v_a. That is, $a > c > d > b$, and, consequently, $(v_a, v_b), (v_d, v_b) \in E(G(S_d))$. However, v_c dominates v_d in $G(S_d)$. Hence $(v_c, v_b) \in E(G)$, a contradiction.

Finally, let (v_a, v_c) be directed from v_a to v_c in $\vec{T}$. Since $a > b$, P contains some vertex v_i, $i \neq a, b$ having in-degree > 1. Such v_i satisfies $i > a$ and $(v_a, v_b), (v_i, v_b) \in E(G(S_a))$. However, v_c dominates v_a in $G(S_a)$. Hence $(v_c, v_b) \in E(G)$, again a contradiction. □

LEMMA 3.2. *All maximal sequences have the same length.*

Proof. Suppose the contrary. Then there is a graph G with two maximal sequences $S_k = \{v_1, \dots, v_k\}$ and $S'_\ell = \{v'_1, \dots, v'_\ell\}$ satisfying $0 < k < \ell$. Clearly, $0 = k < \ell$ cannot occur, as v'_1 is canonical in $G(S_0) = G$. If $S_k = S'_k$, then S_k is not maximal, and the lemma holds. Otherwise, let q be the smallest integer satisfying $v_q \neq v'_q$. We construct a canonical sequence S''_k in which $S''_{q+1} = S_{q+1}$ and such that S''_k is maximal if and only if S_k is so, leading to a contradiction.

In general, for a vertex v_i, let $\text{dom}(i, S_j)$ denote its set of dominators in $G(S_j)$, not belonging to S_{j-1}. First, we show that, if w, x are vertices $\notin S_{j-1}$ such that $w \in \text{dom}(x, S_i)$, then $w \in \text{dom}(x, S_j)$, for $i < j$. Suppose this domination condition is not true. Then, either $(x, w) \notin E(G(S_j))$ or there exists some vertex y adjacent to x and not to w in $G(S_j)$. However, $w \in \text{dom}(x, S_i)$. Then $(x, w) \in E(G(S_i))$. Since $i < j$ and $x, w \notin S_{j-1}$, it follows that $(x, w) \in E(G(S_j))$. For the second alternative, suppose that (y, x) is an edge of $G(S_j)$. Then (y, x) is an edge of $G(S_\ell)$, $\ell < j$. That is, y is incident in $G(S_\ell)$ to at least two vertices $w, x \notin S_{j-1}$. This implies (y, w) to be an edge of $G(S_j)$, a contradiction. Hence w dominates x in S_j, and the assertion is proved.

Examine the unmatched vertex v'_q. The following can occur.

Case 1. $v'_q \in S_k$.

Then $v'_q = v_j$, for some $j > q$.

Case 1.1. $\text{dom}(v_j, S_q) - S_{j-1} \neq \emptyset$.

Let $w \in \text{dom}(v_j, S_q) - S_{j-1}$. Let S''_k be the sequence obtained from S_k by moving v_j to the qth position, while maintaining the relative ordering of the remaining vertices. We show that S''_k is canonical. Let $z \in \text{dom}(v_i, S_i)$, $q \leq i < j$. If $z \neq v_j$, it follows that z also dominates $v_i = v''_{i+1}$ in $G(S''_{i+1})$. Consider $z = v_j$. By the above domination-preserving condition, $w \in \text{dom}(v_j, S_j)$. If there is v_ℓ, $q+1 \leq \ell \leq j-1$, such that $v_j \in \text{dom}(v_\ell, S_\ell)$, then, because $w \in \text{dom}(v_j, S_\ell)$, it follows $w \in \text{dom}(v_\ell, S_\ell)$. Hence S''_k is canonical. In addition, the vertices of S''_k and S_k are the same; i.e., S''_k is maximal. However, $S''_{q+1} = S'_{q+1}$, while $S_{q+1} \neq S'_{q+1}$.

Case 1.2. $\text{dom}(v_j, S_q) - S_{j-1} = \emptyset$.

By the domination condition, there exists some $v_i \in \{v_q, \dots, v_{j-1}\} \cap \text{dom}(v_j, S_i)$. Moreover, choose v_i so that no v_ℓ, $\ell = i+1, \dots, j-1$ dominates v_j in $G(S_\ell)$. Let $S^\star_k$ be the sequence obtained by interchanging the positions of v_i and v_j in S_k. Let $z \in \text{dom}(v_i, S_i)$. If $z = v_j$, then $N_{G(S_i)}[v_i] = N_{G(S_i)}[v_j]$, and $S^\star_k$ is canonical. If $z \neq v_j$, then $z \neq v_\ell$, $\ell = i+1, \dots, j-1$; otherwise, v_ℓ dominates v_j in $G(S_\ell)$, a contradiction. Hence $v_j \in \text{dom}(v_p, S_p)$ implies that $z \in \text{dom}(v_p, S_p)$, $i \leq p \leq j$. Therefore, $S^\star_k$ is canonical and maximal. Then Case 1.1 applies.

Case 2. $v'_q \notin S_k$.

Let $w \in \text{dom}(v'_q, S_q)$. Then $w = v_i$, for some $i \leq k$; otherwise, S_k is not maximal, by the domination property. Moreover, choose v_i such that no v_ℓ, $\ell > i$ dominates v'_q in $G(S_\ell)$. Let $z \in \text{dom}(v_i, S_i)$. Suppose that $z \neq v'_q$. It follows that $z \in \text{dom}(v'_q, S_i)$. If $z \notin S_k$, then $z \in \text{dom}(v'_q, S_k)$, and S_k is not maximal. Then $z \in S_k$. Now, however, $z = v_\ell$ for some $\ell > i$ and $v_\ell \in \text{dom}(v'_q, S_\ell)$, a contradiction. Hence $z = v'_q$. In this case, $N_{G(S_i)}[v_q{}'] = N_{G(S_i)}[v_i]$. Replace v_i by v'_q in S_k. The new sequence so obtained is also canonical and maximal. Then Case 1 applies. □

Theorem 2 and Lemma 1 lead to a greedy algorithm for recognizing expanded trees. Construct a maximal canonical sequence S_k of vertices $v_1, \dots, v_k$ of the graph G. Clearly, G is an expanded tree if and only if $k = n$. For $i < n$, each v_i can be arbitrarily chosen among the dominated vertices in $G(S_i)$, if existing. The algorithm terminates within $O(n^2m)$ steps. A canonical tree T can be obtained as a by-product: For $i < n$, include in $E(T)$ the edge (v_i, w), where w is the dominator of v_i in $G(S_i)$.

Acknowledgments. A referee has pointed out that expanded trees were previously investigated by F. F. Dragan in his Ph.D. thesis (*Centers of Graphs and the Helly Property*, Minsk, 1989, in Russian) under the name *HT-graphs*, where the equivalence (ii)⇔(iii) of Theorem 1 was already established. We are grateful to the referees for the suggestions that improved this article. Particulalrly, H. Bandelt provided new proofs of Corollary 1 and part of Theorem 1, which have been adopted.

REFERENCES

[1] H. Bandelt and E. Pesch, *Dismantling absolute retracts of reflexive graphs*, European J. Combin., 10 (1989), pp. 211–220.

[2] H. Bandelt and E. Prisner, *Clique graphs and Helly graphs*, J. Combin. Theory Ser. B, 51 (1991), pp. 34–45.

[3] P. Buneman, *A characterization of rigid circuit graphs*, Discrete Math., 9 (1974), pp. 205–212.

[4] B. Chen and K. Lih, *Diameters of iterated graphs of chordal graphs*, J. Graph Theory, 14 (1990), pp. 391–396.

[5] F. Gavril, *The intersection graph of subtrees of a tree are exactly the chordal graphs*, J. Combin. Theory Ser. B, 16 (1974), pp. 47–56.

[6] C. L. Monma and V. K. Wei, *Intersection graphs of paths in a tree*, J. Combin. Theory Ser. B, 41 (1986), pp. 141–181.

[7] R. Nowakovski and I. Rival, *The smallest graph variety containing all paths*, Discrete Math., 43 (1983), pp. 223–234.

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

[9] E. Prisner, *Convergence of iterated clique graphs*, Discrete Math., 103 (1992), pp. 199–207.

[10] A. Quilliot, *On the Helly property working as a compactness criterion on graphs*, J. Combin. Theory Ser. A, (1985), pp. 186–193.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 337–343, August 1994

ON SYSTEMATIC CODES OBTAINED FROM LINEAR CODES OVER $GF(q^2)$*

C. MOUAHA[†]

Abstract. A characterization is given of systematic codes over $GF(q)$ that are q-ary images of linear codes over $GF(q^2)$.

Key words. q-ary image, self-dual code, cyclic code

AMS subject classification. 94B05

Introduction. Linear codes that are q-ary images were studied by MacWilliams [4], Mouaha [5], [6], and Wolfmann [8], among others; Beenker [1] examined double circulant codes, which are obtained from linear codes over $GF(4)$ and $GF(9)$. Goldberg [2] reconstructed the ternary Golay code.

In this paper, we give a characterization of systematic codes over $GF(q)$ that are q-ary images of linear codes over $GF(q^2)$. The characterization utilizes the concept of minimal polynomials of elements of $GF(q^2)$.

The plan of this paper is as follows. Section 1 gives some general properties in the study of q-ary images. Section 2 characterizes systematic codes over $GF(q)$ obtained from linear codes over $GF(q^2)$. Section 3 deals with cyclic codes. Section 4 studies systematic self-dual codes that are q-ary images over $GF(q^2)$. Section 5 constructs some linear codes that map onto linear codes over $GF(q^2)$.

1. Preliminaries. Let $GF(q)$ be the finite field of q elements and let $GF(q^m)$ be the extension of degree m of $GF(q)$. Let $B = (b_1, \ldots, b_m)$ be a basis of $GF(q^m)$ over $GF(q)$ and let h_B be the mapping from $GF(q^m)^n$ to $GF(q)^{mn}$ defined by $h_B(\underline{c}) = (\underline{c}_1, \ldots, \underline{c}_m)$, for all $\underline{c} = (c_1, \ldots, c_n) = \sum_{i=1}^{m} \underline{c}_i b_i \in GF(q^m)^n$, where, for all $1 \leq i \leq n$, $c_i = \sum_{j=1}^{m} c_{ij} b_j$ and, for all $1 \leq j \leq m$, $\underline{c}_j = (c_{1j}, \ldots, c_{nj})$. It is clear that h_B is an isomorphism from $GF(q^m)^n$ to $GF(q)^{mn}$, and, consequently, if C is a linear code over $GF(q^m)$, $h_B(C)$ is a linear code over $GF(q)$. $h_B(C)$ is called the q-ary image of C with respect to B.

Remark 1.1. A monomial matrix is a matrix with exactly one nonzero entry in each row and column. Two linear codes C_1 and C_2 both of length n over $GF(q)$ are called equivalent if there is an $n \times n$ monomial matrix Δ over $GF(q)$ such that $C_2 = \{\underline{c}\Delta, \underline{c} \in C_1\}$.

The terminology in this paper is the same as the one used in [5] and [6]. It is important to note that h_B and the map d_B studied in [5] and [6] transform a linear code to equivalent codes.

We have the following obvious properties.

Property 1.1. If $\{\underline{u}_i, i = 1, \ldots, k\}$ is a basis of a linear code C over $GF(q^m)$ and $B = (b_1, \ldots, b_m)$ is a basis of $GF(q^m)$ over $GF(q)$, then $\{h_B(b_i\underline{u}_j), 1 \leq i \leq m, 1 \leq j \leq k\}$ is a basis of $h_B(C)$ over $GF(q)$.

Property 1.2. Let C be a linear code over $GF(q^m)$ and let B a basis of $GF(q^m)$ over $GF(q)$. Then, for all nonzero λ of $GF(q^m)$, $h_{\lambda B}(C) = h_B(C)$.

Property 1.3. If C is an (n, k, δ) linear code over $GF(q^m)$ and B a basis of $GF(q^m)$ over $GF(q)$, then the minimum distance δ' of $h_B(C)$ satisfies $\delta \leq \delta'$.

* Received by the editors September 24, 1992; accepted for publication (in revised form) April 15, 1993.

[†] Département de Mathématiques, Ecole Normale Superieure de Yaoundé, B.P. 47 Yaoundé, Cameroun.

2. Characterization of systematic codes over $GF(q)$ obtained from linear codes over $GF(q^m)$. A systematic code is an (n, k) linear code with generator matrix

$$[I_k \quad A],$$

where I_k is the $k \times k$ identity matrix. It is well known that every (n, k) linear code $k \geq 1$ is equivalent to a systematic code.

In this section and in the following ones, m_λ denotes the minimal polynomial of $\lambda \in GF(q^2)$. Let $B = (1, \lambda)$ be a basis of $GF(q^2)$ over $GF(q)$ and let A be an $n \times n$ matrix over $GF(q)$ such that $m_\lambda(A) = 0$. Let $T(\lambda, A) = \{a_0 I_n + a_1 A,\ a_0, a_1 \in GF(q)\}$ and let θ_λ be the mapping from $GF(q^2)$ to $T(\lambda, A)$ defined by $\theta_\lambda(a_0 + a_1\lambda) = a_0 I_n + a_1 A$, for all a_0 and a_1 in $GF(q)$. Then we have the following result.

Property 2.1. θ_λ is a field isomorphism from $GF(q^2)$ to $T(\lambda, A)$.

Proof. θ_λ is obviously a ring epimorphism. Let $(a_0, a_1) \in GF(q)^2$ such that $\theta_\lambda(a_0 + a_1\lambda) = 0$. Since $A \notin \{0, I_n\}$, then there is $(i, j) \in \{1, \ldots, n\}^2$, $i \neq j$ such that $a_{ij} \neq 0$, where

$$A = (a_{ij})_{1 \leq i,j \leq n}.$$

Therefore, $a_1 = 0$ and $a_0 = 0$.

The following theorem is the fundamental result of this paper.

THEOREM 2.1. *Let D be a $(4k, 2k)$ linear code over $GF(q)$ with generator matrix*

$$G = [I_{2k} \quad A]$$

and let $B = (1, \lambda)$ be a basis of $GF(q^2)$ over $GF(q)$. Then $h_B^{-1}(D)$ is a linear code over $GF(q^2)$ if and only if $m_{-1/\lambda}(A) = 0$.

Proof. Assume that $C = h_B^{-1}(D)$ is a linear code over $GF(q^2)$. Then, for all $\underline{c} = \underline{c}_0 + \underline{c}_1\lambda \in C$, $\lambda\underline{c} = r\underline{c}_1 + \lambda(\underline{c}_0 + s\underline{c}_1) \in C$, where $\lambda^2 = r + s\lambda$, for some r, s in $GF(q)$. Therefore, the mapping ψ from D to D defined by $\psi((\underline{c}_0, \underline{c}_1)) = (r\underline{c}_1, \underline{c}_0 + s\underline{c}_1)$ is an automorphism of D. Thus,

$$\psi(G) = [rA \quad I_{2k} + sA]$$

is also a generator matrix of D. Since D is a systematic code, then

$$(rA)^{-1}(I_{2k} + A) = A,$$

that is, $m_{-1/\lambda}(A) = 0$.

Conversely, assume that $m_{-1/\lambda}(A) = 0$. Then $rA^2 = I_{2k} + sA$. Let $\underline{v} \in C$, $C = h_B^{-1}(D)$. Then there is $\underline{u} \in GF(q)^{2k}$ such that $h_B(\underline{v}) = (\underline{u}, \underline{u}A)$, and so $\underline{v} = \underline{u} + \lambda\underline{u}A$. Since A is an invertible matrix, rAG is also a generator matrix of D. Thus $\underline{u}(rAG) = h_B(a\underline{v}) \in D$. Therefore C is a linear code over $GF(q^2)$.

In the following, $D(k, A)$ denotes the $(4k, 2k)$ linear code with generator matrix

$$[I_{2k} \quad A].$$

COROLLARY 2.1. *Let $\lambda \in GF(q^2) - GF(q)$ and let A be a $2k \times 2k$ matrix over $GF(q)$ such that $m_\lambda(A) = 0$. Then, for all $U \in T(\lambda, A)$ such that $\theta_\lambda^{-1}(U) \notin GF(q)$, there is a basis B of $GF(q^2)$ over $GF(q)$ such that $h_B^{-1}(D(k, U))$ is a linear code over $GF(q^2)$.*

Proof. Let $U \in T(\lambda, A)$ such that $\theta_\lambda^{-1}(U) \notin GF(q)$. Let γ be a root of the minimal polynomial of $\theta_\lambda^{-1}(U)$ and

$$B = \left(1, -\frac{1}{\gamma}\right).$$

Then, by Theorem 2.1, $h_B^{-1}(D(k, U))$ is a linear code over $GF(q^2)$.

The following result gives a necessary and sufficient condition under which a q-ary image of a systematic code over $GF(q^2)$ is systematic.

THEOREM 2.2. *Let C be a $(2k, k)$ linear code over $GF(q^2)$ with generator matrix*

$$G = [I_k \quad A]$$

and $B = (1, \lambda)$ a basis of $GF(q^2)$ over $GF(q)$. Let $A = A_1 + \lambda A_2$. Then $h_B(C)$ is a systematic code if and only if A_2 is an invertible matrix.

Proof. Let $\underline{g}_i$ be the ith rowvector of G, $1 \le i \le k$. Then the result follows from Property 1.1 because $h_B(C)$ has generator matrix

$$\begin{bmatrix} h_B(\underline{g}_1) \\ \vdots \\ h_B(\underline{g}_k) \\ h_B(\lambda \underline{g}_1) \\ \vdots \\ h_B(\lambda \underline{g}_k) \end{bmatrix} = \begin{bmatrix} I_k & A_1 & 0 & A_2 \\ 0 & rA_2 & I_k & A_1 + sA_2 \end{bmatrix},$$

where $\lambda^2 = r + s\lambda$, for some $(r, s) \in GF(q)^2$.

Remark 2.1. Let C be a $(2k, k)$ linear code over $GF(q^2)$ with generator matrix

$$[I_k \quad A].$$

Let $A = A_1 + \lambda A_2$, where $B = (1, \lambda)$ is a basis of $GF(q^2)$ over $GF(q)$. Then $h_B(C)$ is equivalent to the systematic code over $GF(q)$ with generator matrix

$$\begin{bmatrix} I_k & 0 & A_1 & A_2 \\ 0 & I_k & rA_2 & A_1 + sA_2 \end{bmatrix},$$

where $\lambda^2 = r + s\lambda$, for some $(r, s) \in GF(q)^2$.

Conversely, let D be a $(4k, 2k)$ linear code over $GF(q)$ with generator matrix of the form

$$\begin{bmatrix} I_k & 0 & U_1 & U_2 \\ 0 & I_k & rU_2 & U_1 + sU_2 \end{bmatrix},$$

where, for all $1 \le i \le 2$, U_i is a $k \times k$ matrix over $GF(q)$. If $X^2 - sX - r$ is an irreducible polynomial over $GF(q)$, then $h_B(C_1)$ is equivalent to D, where C_1 is the linear code over $GF(q^2)$ with generator matrix

$$[I_k \quad U],$$

$U = U_1 + \lambda U_2$, $\lambda^2 = r + s\lambda$, and $B = (1, \lambda)$.

3. The case of cyclic codes. Cyclic codes are an important class of systematic codes. Here we prove that, if $D(k, A)$ is a cyclic code over $GF(q)$, then, for all $\lambda \in GF(q^2) - GF(q)$, $m_\lambda(A) \neq 0$. T denotes the shift operator of $GF(q)^n$.

THEOREM 3.1. *Let C be an (n, k) linear code over $GF(q^2)$ and let $B = (b_0, b_1)$ be a basis of $GF(q^2)$ over $GF(q)$. Then $T^n(h_B(C)) = h_B(C)$ if and only if C is generated by $C_0 = C \cap GF(q)^n$.*

Proof. Assume that $T^n(h_B(C)) = h_B(C)$. Let $\underline{c} = \underline{c}_0 b_0 + \underline{c}_1 b_1 \in C$. Then $(\underline{c}_0, \underline{c}_1) \in h_B(C)$ and $(\underline{c}_1, \underline{c}_0) \in h_B(C)$. Therefore $\underline{c}_0 + \underline{c}_1 \in C$. Since $\underline{c} - b_0(\underline{c}_0 + \underline{c}_1)$ and $\underline{c} - b_1(\underline{c}_0 + \underline{c}_1)$ are codewords of C, then $\underline{c}_0 \in C$ and $\underline{c}_1 \in C$.

Conversely, assume that C is generated by C_0. Then $h_B(C) = C_0^2$.

If C is an (n, k) linear code over $GF(q^2)$, $1 \le k \le n - 1$, then $h_B(C)$ cannot be a cyclic code over $GF(q)$, as we show in the next theorem.

THEOREM 3.2. *Let C be an (n, k) linear code over $GF(q^2)$, $1 \leq k \leq n - 1$. Then, for all basis B of $GF(q^2)$ over $GF(q)$, $h_B(C)$ is not a cyclic code.*

Proof. By Theorem 3.1, it is sufficient to prove that, if D is an (m, k') linear code over $GF(q)$, $1 \leq k' \leq m - 1$, then D^2 is not a cyclic code. Let D be a (m, k') linear code over $GF(q)$, $1 \leq k' \leq m - 1$. Let $\underline{u} = (u_1, \dots, u_m) \in D$, $\underline{u} \neq 0$. Then $\underline{v} = (\underline{u}, 0, \dots, 0) \in D^2$. Assume that D^2 is a cyclic code. Then $T(\underline{v}) = (0, u_1, \dots, u_m, 0, \dots, 0)$; so $\underline{x} = (u_m, 0, \dots, 0) \in D$. Without loss of generality, assume that $u_m \neq 0$. Since $\underline{w} = (\underline{x}, 0, \dots, 0) \in D^2$, then $\{T^i(\underline{w}), 0 \leq i \leq 2m - 1\}$ is a basis of D^2, and so $D^2 = GF(q)^{2m}$. This contradicts the fact that $1 \leq k' \leq m - 1$.

Since every (n, k) cyclic code, $1 \leq k \leq n - 1$, is a systematic code, we have the following corollary.

COROLLARY 3.1. *Let D be a $(4k, 2k)$ cyclic code over $GF(q)$ and let*

$$[I_{2k} \quad A]$$

be its generator matrix. Then, for all $\lambda \in GF(q^2) - GF(q)$, $m_\lambda(A) \neq 0$.

Proof. The result follows from Theorems 2.1 and 3.2.

4. Systematic self-dual q-ary images of linear codes over $GF(q^2)$.

Self-dual codes are an important class of linear codes because of their algebraic and combinatorial properties. In this section, we study systematic self-dual codes that can be constructed as q-ary images over $GF(q^2)$.

We need the following two lemmas.

LEMMA 4.1. *There is $\lambda \in GF(q^2) - GF(q)$ such that $\lambda^{q+1} = -1$.*

Proof. Let α be a primitive element of $GF(q^2)$. Since α^{q+1} is a primitive element of $GF(q)$, there is $i \in \{1, \dots, q - 1\}$ such $\alpha^{i(q+1)} = -1$. Obviously, $\alpha^i \notin GF(q)$.

LEMMA 4.2. *There is $\lambda \in GF(q^2) - GF(q)$ such that $\lambda^2 = -1$ if and only if $q \equiv 3 (\mathrm{mod}\ 4)$.*

Proof. Assume that there is $\lambda \in GF(q^2) - GF(q)$ such that $\lambda^2 = -1$. Then q is an odd number. So $\lambda^{q^2-1} = (-1)^{(q^2-1)/2} = 1$. Therefore $q \equiv 3 (\mathrm{mod}\ 4)$.

Conversely, assume that $q \equiv 3\ (\mathrm{mod}\ 4)$. Let α be a primitive element of $GF(q^2)$. Then there is $i \in \{1, \dots, q - 2\}$ such that $\alpha^{i(q+1)} = -1$. Thus we have $\alpha^{iq(q+1)/2} = -\alpha^{i(q+1)/2}$, and so $\alpha^{i(q+1)/2} \notin GF(q)$.

$C^\perp$ denotes the orthogonal of the linear code C.

We have the following theorem.

THEOREM 4.1. *Let $B = (1, \lambda)$ be a basis of $GF(q^2)$ over $GF(q)$. Then the following statements are equivalent:*

1) $\lambda^{q+1} = -1$,

2) $h_B(C^\perp) = h_B(C)^\perp$ *for all linear code C over $GF(q^2)$.*

Proof. Assume that $\lambda^{q+1} = -1$. Then there is $s \in GF(q)$ such that $\lambda^2 = 1 + s\lambda$. Let C be a linear code over $GF(q^2)$, $\underline{c} \in C$, and $\underline{c}' \in C^\perp$. Then $\langle h_B(\underline{c}), h_B(\underline{c}')\rangle = 0$, (where $\langle \underline{u}, \underline{v}\rangle$ denotes the usual inner product of $\underline{u}$ and $\underline{v}$), and so $h_B(\underline{c}') \in h_B(C)^\perp$. Therefore the result follows from the fact that $h_B(C)^\perp$ and $h_B(C^\perp)$ have the same dimension over $GF(q)$.

Conversely, assume that $h_B(C^\perp) = h_B(C)^\perp$ for all linear code C over $GF(q^2)$. Let $\lambda^2 = r + s\lambda$ for some $(r, s) \in GF(q)^2$. Let C be the linear code generated by $(1, \lambda)$. Since $h_B(C^\perp) = h_B(C)^\perp$, then $\langle h_B((\lambda, \lambda^2)), h_B((\lambda, -1))\rangle = 0$. Therefore $r = 1$.

The following theorem characterizes systematic self-dual codes that are q-ary images over $GF(q^2)$.

THEOREM 4.2. *Let $B = (1, \lambda)$ be a basis of $GF(q^2)$ over $GF(q)$, $\lambda^2 = r + s\lambda$ for some $(r, s) \in GF(q)^2$. Let D be a $(4k, 2k)$ self-dual code over $GF(q)$ with generator matrix*

$$[I_{2k} \quad A].$$

Then $h_B^{-1}(D)$ is a linear code over $GF(q^2)$ if and only if $rA + A^t = sI_{2k}$.

Proof. Assume that $h_B^{-1}(D)$ is a linear code over $GF(q^2)$. Then, by Theorem 2.1, $rA^2 = I_{2k} + sA$. Since $AA^t = -I_{2k}$, we obtain $rA + A^t = sI_{2k}$.

Conversely, assume that $rA + A^t = sI_{2k}$. Since D is a self-dual code, then $rA^2 - I_{2k} = sA$. The result follows from Theorem 2.1.

The following corollaries are obvious.

COROLLARY 4.1. *Let $B = (1, \lambda)$ be a basis of $GF(q^2)$ over $GF(q)$, $q \equiv 0(\text{mod } 2)$, such that $\lambda^{q+1} = 1$. Let D be a $(4k, 2k)$ systematic self-dual code over $GF(q)$. Then $h_B^{-1}(D)$ is not a linear code over $GF(q^2)$.*

COROLLARY 4.2. *Let $B = (1, \lambda)$ be a basis of $GF(q^2)$ over $GF(q)$, $q \equiv 3(\text{mod } 4)$, such that $\lambda^2 = -1$. Let D be a $(4k, 2k)$ self-dual code over $GF(q)$ with generator matrix*

$$[I_{2k} \quad A].$$

Then $h_B^{-1}(D)$ is a linear code over $GF(q)$ if and only if $A = A^t$.

Example 4.1. Let

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix}$$

be a matrix over $GF(3)$. Then $A^2 = A + I_2$ and $AA^t = -I_2$. Therefore the linear code D with generator matrix

$$[I_2 \quad A]$$

is self-dual, and $h_B^{-1}(D)$ is a self-dual code over $GF(9)$, $B = (1, \alpha^3)$, $\alpha^2 = 1 + \alpha$.

5. Construction of q-ary images of some linear codes over $GF(q^2)$. Let B be a basis of $GF(q^m)$ over $GF(q)$, $a \in GF(q^m)$ and let μ_a be the mapping from $GF(q^m)$ to $GF(q^m)$ defined by $\mu_a(x) = ax$. Obviously if $a \neq 0$, μ_a is an automorphism of $GF(q^m)$ over $GF(q)$. Let M_a^t be the matrix of μ_a with respect to B and let $M(B) = \{M_x, x \in GF(q^m)\}$. Then $M(B)$ is a finite field of q^m elements. Let η_B be the mapping of $GF(q^m)$ onto $M(B)$ by $\eta_B(x) = M_x$. Then η_B is a field isomorphism from $GF(q^m)$ to $M(B)$. In this paper, every finite field of $m \times m$ matrices over $GF(q)$ is a matrix representation of $GF(q^m)$.

THEOREM 5.1. *Let $f \in GF(q)[x]$ be an irreducible polynomial of degree 2 over $GF(q)$ and let $A = (a_{ij})_{1 \leq i,j \leq 2t}$ be a matrix over $GF(q^{2m})$ such that $f(A) = 0$. Let B be a basis of $GF(q^{2m})$ over $GF(q)$ and let D be the $(4mt, 2mt)$ linear code over $GF(q)$ with generator matrix*

$$[I_{2mt} \quad (\eta_B(a_{ij}))_{1 \leq i,j \leq 2t}].$$

Then there is a basis B_1 of $GF(q^2)$ over $GF(q)$ such that $h_{B_1}^{-1}(D)$ is a linear code over $GF(q^2)$.

Proof. Set

$$U = (\eta_B(a_{ij}))_{1 \leq i,j \leq 2t}.$$

Since $f(A) = 0$ and η_B is a field isomorphism from $GF(q^m)$ to $M(B)$, then $f(U) = 0$. Thus the result follows from Theorem 2.1.

COROLLARY 5.1. *Let B be a basis of $GF(q^{2m})$ over $GF(q)$. Then, for all $x \in GF(q^2) - GF(q)$, there is a basis B_1 of $GF(q^2)$ over $GF(q)$ such that $h_{B_1}^{-1}(D_x)$ is a linear code over $GF(q^2)$, where D_x is the linear code over $GF(q)$ with generator matrix*

$$[I_{2m} \quad \eta_B(x)].$$

Proof. Let $x \in GF(q^2) - GF(q)$. Then $m_x(\eta_B(x)) = 0$. The result follows from Theorem 5.1.

DEFINITION 5.1. *A basis B of $GF(q^m)$ over $GF(q)$ is a symmetric basis if, for all $A \in M(B)$, $A = A^t$.*

In [7] it is established that every finite field $GF(q^m)$ has a symmetric basis over $GF(q)$. The following result is a direct consequence of Theorem 3.2 of [6].

THEOREM 5.2. *Let B be a basis of $GF(q^m)$ over $GF(q)$. Then the following statements are equivalent:*

1) *B is a symmetric basis,*

2) *$h_B(C^\perp) = h_B(C)^\perp$ for every linear code C over $GF(q^m)$.*

Again, we have the following theorem.

THEOREM 5.3. *Let $q \equiv 3(\text{mod } 4)$ and let $A = (a_{ij})_{1 \leq i,j \leq 2t}$ be a symmetric matrix over $GF(q^{2m})$ such that $A^2 = -I_{2t}$. Then, for all symmetric basis B of $GF(q^{2m})$ over $GF(q)$, we have the following facts:*

1) *The linear code D with generator matrix*

$$[I_{2mt} \quad (\eta_B(a_{ij}))_{1 \leq i,j \leq 2t}]$$

is a self-dual code;

2) *There is a basis B_1 of $GF(q^2)$ over $GF(q)$ such that $h_{B_1}^{-1}(D)$ is a linear code over $GF(q^2)$.*

Proof. 1) Since A is a symmetric matrix and B a symmetric basis, then

$$U = (\eta_B(a_{ij}))_{1 \leq i,j \leq 2t}$$

is symmetric and $U^2 = -I_{2mt}$.

2) Since $U^2 = -I_{2mt}$, the result follows from Theorem 5.1.

Example 5.1. Let α be a primitive element of $GF(9)$ and let

$$A = \begin{pmatrix} \alpha & \alpha^2 & \alpha^3 \\ \alpha^2 & 1 & \alpha^2 \\ \alpha^3 & \alpha^2 & 1 \end{pmatrix}.$$

Then A is a symmetric matrix and $A^2 = -I_3$. Let B be a symmetric basis of $GF(3^{2m})$ over $GF(3)$ and let $U = \eta_B(\alpha)$. Then the linear code D_m with generator matrix

$$[I_{6m} \quad M],$$

where

$$M = \begin{pmatrix} U & U^2 & U^3 \\ U^2 & I_{2m} & U^2 \\ U^3 & U^2 & I_{2m} \end{pmatrix}$$

is a $(12m, 6m, \delta_m)$ self-dual code over $GF(3)$ with $\delta_m \geq 6$. Let $B_1 = (1, \gamma)$ be a basis of $GF(9)$ over $GF(3)$ such that $\gamma^2 = -1$. Then $h_{B_1}^{-1}(D_m)$ is a linear code over $GF(9)$.

We need a result that can be found in [7], stated here in the form of the following lemma.

LEMMA 5.1. *R is a matrix representation of $GF(q^m)$ if and only if there exists a basis B of $GF(q^m)$ over $GF(q)$ such that $R = M(B)$.*

We end this section with the following theorem.

THEOREM 5.4. *Let D be a $(4m, 2m)$ linear code over $GF(q)$ with generator matrix*

$$[I_{2m} \quad A].$$

If there is λ in $GF(q^2) - GF(q)$ such that $m_\lambda(A) = 0$, then there exist a linear code C over $GF(q^{2m})$ and a basis B of $GF(q^{2m})$ over $GF(q)$ such that $h_B(C)$ and D are equivalent codes.

Proof. Assume that there is $\lambda \in GF(q^2) - GF(q)$ such that $m_\lambda(A) = 0$. Since $T(\lambda, A)$ is a field of q^2 elements and $GF(q^2)$ is a subfield of $GF(q^{2m})$, then, by Lemma 5.1, there is a basis B of $GF(q^{2m})$ over $GF(q)$ such that $T(\lambda, A) \subset M(B)$. Let $v \in GF(q^{2m})$ such that $\eta_B(v) = A$ and let C be the linear code over $GF(q^{2m})$ generated by $(1, v)$. Then $h_B(C)$ and D are equivalent codes.

6. Conclusion. This paper has shown interesting algebraic properties on systematic q-ary images of linear codes over $GF(q^2)$. We hope that, with the help of this work, other important linear codes can be either constructed or reconstructed and new properties discovered about them.

Acknowledgments. The author is grateful to the anonymous referees for their useful suggestions.

REFERENCES

[1] G. F. M. BEENKER, *On Double Circulant Codes*, T.H. Report 80-W S K-04, July 1980.

[2] D. Y. GOLDBERG, *Reconstructing the ternary Golay code*, J. Combin. Theory, A42 (1986), pp. 296–299.

[3] M. KARLIN AND F. J. MACWILLIAMS, *Quadratic residue codes over GF(4) and their binary images*, IEEE Internat. Sympos. on Inform. Theory, Asilomar, CA, 1972.

[4] F. J. MACWILLIAMS, *On binary cyclic codes which are also cyclic over $GF(2^s)$*, SIAM J. Appl. Math., 19 (1970), pp. 75–95.

[5] C. MOUAHA, *On cyclic codes which are q-ary images of linear codes*, Appl. Alg. Engrg., Comm. Comput., 2 (1992), pp. 163–170.

[6] ———, *On q-ary images of self-dual codes*, Appl. Alg. Engrg., Comm. Comput., 3 (1992), pp. 311–319.

[7] G. SEROUSSI AND A. LEMPEL, *On symmetric representations of finite fields*, SIAM J. Alg. Discrete Math., 4 (1983), pp. 14–21.

[8] J. WOLFMANN, *Différents aspects de la démultiplication des codes*, Traitement du signal, 1 (1984).

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 344-358, August 1994

COMPOSITIONS OF GRAPHS AND POLYHEDRA I: BALANCED INDUCED SUBGRAPHS AND ACYCLIC SUBGRAPHS*

FRANCISCO BARAHONA[†] AND ALI RIDHA MAHJOUB[‡]

Abstract. Let $P(G)$ be the balanced induced subgraph polytope of G. If G has a two-node cutset, then G decomposes into G_1 and G_2. It is shown that $P(G)$ can be obtained as a projection of a polytope defined by a system of inequalities that decomposes into two pieces associated with G_1 and G_2. The problem max cx, $x \in P(G)$ is decomposed in the same way. This is applied to series-parallel graphs to show that, in this case, $P(G)$ is a projection of a polytope defined by a system with $O(n)$ inequalities and $O(n)$ variables, where n is the number of nodes in G. Also for this class of graphs, an algorithm is given that finds a maximum weighted balanced induced subgraph in $O(n \log n)$ time. This approach is also used to obtain composition of facets of $P(G)$. Analogous results are presented for acyclic induced subgraphs.

Key words. polyhedral combinatorics, composition of polyhedra, balanced subgraphs, acyclic subgraphs, compact systems

AMS subject classifications. 05C85, 90C27

1. Introduction. Given a graph G, let $P(G)$ be a polytope associated with G. If G has a one- or two-node cutset, then G decomposes into G_1 and G_2. We study a technique to derive $P(G)$, provided that we know two polytopes related to G_1 and G_2. We use the same ideas to decompose the problem

$$\text{Maximize } cx,$$
$$x \in P(G)$$

into two optimization problems related to G_1 and G_2. Similar compositions of polyhedra have been studied in [8], [3], [10], [5], and [9]. First, we study the polytopes of balanced induced subgraphs and acyclic induced subgraphs. In a subsequent paper, we apply a simplification of this technique to the stable set polytope.

The graphs we consider are finite, undirected, and may have multiple edges. We denote a graph by $G = (V, E)$, where V is the *node set* and E is the *edge set* of G. If $W \subseteq V$, then $E(W)$ denotes the set of all edges of G with both endnodes in W, and the graph $H = (W, E(W))$ is the subgraph of G induced by W.

A signed graph $G = (V, E)$ is a graph whose edges are labeled positive or negative [11]. A positive (negative) edge $\{u, v\}$ is denoted by $\{u, v, +\}$, $(\{u, v, -\})$. A signed graph is said to be *balanced* if the set of negative edges form a cut, that is, if the node set V can be partitioned into U and $\bar{U}$ in such a way that $E(U) \cup E(\bar{U})$ is the set of positive edges. Also, a signed graph is balanced if it does not contain a cycle with an odd number of negative edges. Suppose that from a node i we send a signal $s_i \in \{-1, 1\}$ to all the adjacent nodes; if the edge $\{i, j\}$ is positive, then j receives the signal s_i; if $\{i, j\}$ is negative, then j receives $-s_i$. A signed graph is balanced if and only if, when we send a signal from a node, this node receives the same signal in return.

[*] Received by the editors June 4, 1990; accepted for publication (in revised form) May 20, 1993. This research was supported in part by the Natural Sciences and Engineering Research Council, Canada.

[†] IBM T. J. Watson Research Center, Yorktown Heights, New York 10598.

[‡] Department of Statistics, King Saud University, Riyadh, Saudi Arabia. Present address, Laboratoire d'Informatique de Brest (LIBr), Université de Bretagne Occidentale, 6 Avenue Le Gorgeu, 29287 Brest Cedex, France. This author was supported by C. P. Rail.

If $W \subseteq V$ and $H = (W, E(W))$ is balanced, then H is called a *balanced induced subgraph* (BIS) of G. If $W \subseteq V$, let $x^W \in \Re^V$, where $x^W(u) = 1$ if $u \in W$ and where $x^W(u) = 0$ if $u \notin W$; x^W is called the *incidence vector* of W.

The convex hull of incidence vectors of all balanced induced subgraphs of G is denoted by $P(G)$ and called the BIS *polytope* of G, i.e.,

$$P(G) = \text{conv}\,\{x^W \in \Re^V \mid (W, E(W)) \text{ is a BIS of } G\}.$$

Given a signed graph $G = (V, E)$ with node weights $c(v)$ for all $v \in V$, the *maximum* BIS *problem* is to find a BIS $(W, E(W))$ such that $c(W) = \sum \{c(v) : v \in W\}$ is as large as possible.

Every optimum basic solution of the linear program

$$\max cx,$$
$$x \in P(G)$$

is the incidence vector of a maximum weighted BIS of G.

The edge problem of finding a maximum balanced spanning subgraph can be reduced to a max-cut problem [4]. This problem is polynomially solvable for graphs not contractible to K_5 [3] and for toroidal graphs [2].

If $H = (V, F)$ is a graph, then the maximum stable set problem in H can be reduced to a maximum BIS problem in a signed graph $G = (V, E)$ that is obtained by replacing each edge in F by a positive edge and a negative edge. Thus the maximum BIS problem can be viewed as a generalization of the maximum stable set problem. This shows that the maximum BIS problem is NP-hard even for signed planar graphs. When all the edges are negative, a BIS coincides with a bipartite induced subgraph.

The polytope $P(G)$ is full-dimensional. This implies that (up to multiplication by a positive constant) there is a unique nonredundant inequality system $Ax \leq b$ such that $P(G) = \{x \mid Ax \leq b\}$; moreover, there is a natural bijection among the facets of $P(G)$ and the inequalities of that system.

In §2 we show that, if G admits a two-vertex decomposition into G_1 and G_2, then $P(G)$ is a projection of a polytope defined by a system of inequalities that decomposes into pieces associated with G_1 and G_2. In §3 the optimization problem is decomposed in a similar way. In §4 we apply this technique to series-parallel graphs and we show that, in this case, $P(G)$ is a projection of a polytope defined by a system with $O(n)$ inequalities and $O(n)$ variables, where n is the number of nodes in G. Also for this class of graphs, we give an algorithm that finds a maximum BIS in $O(n \log n)$ time. In §5 we use the same approach for finding compositions of facets of $P(G)$. Analogous results about acyclic induced subgraphs are mentioned in §6.

2. Compositions of graphs.

In this section, we derive a system of inequalities that defines a polytope having $P(G)$ as a projection, provided that G is a composition of two graphs and such a system is known for each piece.

The next theorem is a generalization of a result of Chvátal [7] about the stable set polytope.

THEOREM 2.1. *Let $G = (V, E)$ be a signed graph such that there exists node sets V_1 and V_2 with the following properties*:

(i) $V = V_1 \cup V_2$,

(ii) $W = V_1 \cap V_2 \neq \emptyset$,

(iii) *Between each pair $\{i, j\} \subseteq W$, there exists a positive edge and a negative edge in E,*

(iv) *The induced subgraph $(V \backslash W, E(V \backslash W))$ is disconnected.*

If $G_1 = (V_1, E(V_1))$, $G_2 = (V_2, E(V_2))$, then a system of inequalities that defines $P(G)$ is obtained by the juxtaposition of such systems defining $P(G_1)$ and $P(G_2)$.

Proof. Let $Ax \le b$ be the system obtained by juxtaposing both systems. Let $\bar{x}$ be a point in the polyhedron defined by $Ax \le b$. Let $\bar{x}_1$ ($\bar{x}_2$) be the restriction of $\bar{x}$ to the set of indices associated with nodes in V_1 (V_2). Since $\bar{x}_1 \in P(G_1)$, we can write

$$\bar{x}_1 = \sum_{i=1}^{p} \lambda_i y_i,$$

$$\lambda_i \ge 0 \quad \text{for } 1 \le i \le p, \ \sum_{i=1}^{p} \lambda_i = 1,$$

where y_i is an extreme point of $P(G_1)$ for $i = 1, \dots, p$. We can also write

$$\bar{x}_2 = \sum_{i=1}^{k} \alpha_i z_i,$$

$$\alpha_i \ge 0 \quad \text{for } 1 \le i \le k, \ \sum_{i=1}^{k} \alpha_i = 1,$$

where z_i is an extreme point of $P(G_2)$ for $i = 1, \dots, k$.

Let $W = \{w_1, \dots, w_l\}$. Then

$$\bar{x}(w_j) = \sum_i \{\lambda_i \,|\, y_i(w_j) = 1\} = \sum_i \{\alpha_i \,|\, z_i(w_j) = 1\}, \qquad 1 \le j \le l.$$

We can match a vector y_i with $y_i(w_j) = 1$ with a vector z_r with $z_r(w_j) = 1$ for $1 \le j \le l$ and we can match a vector y_i with $y_i(w_j) = 0$ for $1 \le j \le l$ with a vector z_r with $z_r(w_j) = 0$ for $1 \le j \le l$. We obtain an incidence vector of a balanced induced subgraph of G. Thus $\bar{x}$ can be written as a convex combination of incidence vectors of balanced induced subgraphs of G. □

Now we study graphs with a two-vertex cutset. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs such that $V_1 \cap V_2 = \{u, v\}$ and let $G = (V, E)$ be the union of G_1 and G_2, i.e., $V = V_1 \cup V_2$, $E = E_1 \cup E_2$. We cover four cases.

Case 1. There is a positive edge and a negative edge between u and v in G.

Case 2. There is only a positive edge between u and v in G.

Case 3. There is only a negative edge between u and v in G.

Case 4. Nodes u and v are not adjacent in G.

Case 1 is covered by Theorem 2.1; thus we restrict ourselves to the three other cases. In Case 2, we define $\bar{G}_i = (\bar{V}_i, \bar{E}_i)$, $i = 1, 2$ as follows; see Fig. 1:

$$\bar{V}_i = V_i \cup \{w_1, w_2, w_3, w_4\},$$

$$\bar{E}_i = E_i \cup \{\{w_1, u, -\}, \{w_1, v, +\}, \{w_2, u, -\}, \{w_3, u, -\},$$
$$\{w_3, v, -\}, \{w_4, v, -\}, \{w_3, w_4, -\}, \{w_3, w_2, -\}\}.$$

In Case 3, $\bar{G}_i = (\bar{V}_i, \bar{E}_i)$, $i = 1, 2$ are defined as in Case 2, but $\{w_1, v\}$ is labeled negative. In Case 4, $\bar{G}_i = (\bar{V}_i, \bar{E}_i)$, $i = 1, 2$ are defined as follows:

$$\bar{V}_i = V_i \cup \{w_1, w_2, w_3, w_4, w_5\},$$

$$\bar{E}_i = E_i \cup \{\{w_1, u, -\}, \{w_2, u, -\}, \{w_3, u, -\}, \{w_4, u, -\},$$
$$\{w_1, v, -\}, \{w_2, v, +\}, \{v, w_4, -\}, \{v, w_5, -\}, \{w_3, w_4, -\}, \{w_4, w_5, -\}\}.$$

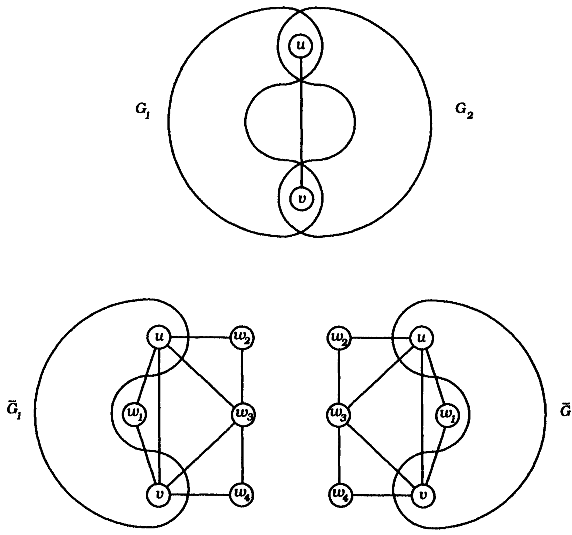


FIG. 1

Our aim is to derive a system of inequalities for $P(G)$ from systems defining $P(\bar{G}_1)$ and $P(\bar{G}_2)$. Let $\bar{G} = (\bar{V}, \bar{E})$ be the graph obtained as union of $\bar{G}_1$ and $\bar{G}_2$, i.e., $\bar{V} = \bar{V}_1 \cup \bar{V}_2$, $\bar{E} = \bar{E}_1 \cup \bar{E}_2$.

In Cases 2 and 3, the inequality

$$(2.1) \qquad \sum_{i=1}^{4} x(w_i) + x(u) + x(v) \leq 4$$

defines a facet $F(\bar{G}_i)$ of $P(\bar{G}_i)$, $i = 1, 2$ and a facet $F(\bar{G})$ of $P(\bar{G})$. In Case 4, the inequality

$$(2.2) \qquad \sum_{i=1}^{5} x(w_i) + x(u) + x(v) \leq 5$$

plays the same role.

The polytope $P(G)$ is the projection of $F(\bar{G})$ along the variables $\{x(w_i)\}$. The next theorem gives us a system defining $F(\bar{G})$.

THEOREM 2.2. *The juxtaposition of a system that defines $F(\bar{G}_1)$ and a system that defines $F(\bar{G}_2)$ gives a system that defines $F(\bar{G})$.*

Proof. Let $Ax \leq b$ be such a system and let $\bar{x}$ be a vector that satisfies it. Let $\bar{x}_1$ ($\bar{x}_2$) be the restriction of $\bar{x}$ to the set of indices associated with nodes in $\bar{V}_1$ ($\bar{V}_2$). Since $\bar{x}_1 \in F(\bar{G}_1)$, we have that

$$\bar{x}_1 = \sum_{i=1}^{p} \lambda_i y_i,$$

$$\lambda_i \geq 0 \quad \text{for } 0 \leq i \leq p, \ \sum_{i=1}^{p} \lambda_i = 1,$$

where y_i is an extreme point of $F(\bar{G}_1)$ for $i = 1, \ldots, p$.

We can also write

$$\bar{x}_2 = \sum_{i=1}^{k} \alpha_i z_i,$$

$$\alpha_i \geq 0 \quad \text{for } 0 \leq i \leq k, \ \sum_{i=1}^{k} \alpha_i = 1,$$

where z_i is an extreme point of $F(\bar{G}_2)$ for $i = 1, \ldots, k$.

Let us study Case 2. For a vector $x \in F(\bar{G})$, let us assume that its last six components are $x(w_i)$, $i = 1, \ldots, 4$, $x(u)$, and $x(v)$. Then, for each vector y_i, its last six components form one column of the matrix

$$M = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

Set $\beta_j = \sum \{\lambda_i \mid y_i$ is associated with the jth column of $M\}$. The vector β satisfies

$$M\beta = \begin{bmatrix} \bar{x}(w_1) \\ \vdots \\ \bar{x}(w_4) \\ \bar{x}(u) \\ \bar{x}(v) \end{bmatrix}.$$

In the same way, we can associate the vectors $\{z_i\}$ with the columns of M.

We can define $\gamma_j = \sum \{\alpha_i \mid z_i$ is associated with the jth column of $M\}$, since γ satisfies

$$M\gamma = \begin{bmatrix} \bar{x}(w_1) \\ \vdots \\ \bar{x}(w_4) \\ \bar{x}(u) \\ \bar{x}(v) \end{bmatrix}$$

and M is nonsingular; we have that $\beta = \gamma$. Hence vectors $\{y_i\}$ can be matched with vectors $\{z_i\}$ in such a way that $\bar{x}$ can be written as a convex combination of extreme points of $F(\bar{G})$.

Cases 3 and 4 are analogous. □

3. Algorithmic aspects of the compositions. In this section, we use the compositions of §2 to obtain a maximum weighted BIS, provided that we have an algorithm to solve the problem in each piece.

Let $G = (V, E)$ be a signed graph and $c : V \rightarrow \Re_+$ a weight function.

First, let us assume that G is the graph of Theorem 2.1, let $W = \{w_1, \ldots, w_l\}$, and let β_i be the maximum weight of a BIS of G_2 that contains w_i for $1 \leq i \leq l$. Let β_0 be the maximum weight of a BIS of G_2 that does not contain any node of W. Let us redefine the weights in G_1 as follows:

$$c'(u) = c(u) \quad \text{if } u \notin W,$$

$$c'(w_i) = \max\{0, \beta_i - \beta_0\} \quad \text{for } 1 \leq i \leq l.$$

Let α be the maximum weight of a BIS of G_1; then the maximum weight of a BIS of G is $\alpha + \beta_0$.

Now let us study Cases 2–4 of §2. Set $W = \bar{V}_1 \cap \bar{V}_2$ and let

$$\sum_{u \in W} x(u) \leq r(W) \tag{3.1}$$

be inequality (2.1) or (2.2). If $\bar{x}$ is the incidence vector of a BIS of G, we can always complete it to an incidence vector of a BIS of $\bar{G}$ that satisfies (3.1).

Let $W = \{v_1, \ldots, v_p\}$. There are p BIS of $(W, \bar{E}(W))$ whose incidence vectors satisfy (3.1); moreover, they are linearly independent. Let $U_1, \ldots, U_p$ be the node sets of them. Let β_i be the maximum weight of a BIS of $\bar{G}_2$ whose node set contains U_i for $1 \leq i \leq p$. The weights for nodes of $W \backslash V_2$ are zero. Let $[\gamma_1 \cdots \gamma_p]$ be the solution of the system of equations

$$[\gamma_1 \cdots \gamma_p][x^{U_1} \cdots x^{U_p}] = [\beta_1 \cdots \beta_p].$$

Let us redefine the weights in $\bar{G}_1$ as follows:

$$c'(u) = c(u) \quad \text{if } u \in V_1 \backslash W,$$

$$c'(v_i) = \gamma_i + M \quad \text{for } 1 \leq i \leq p,$$

where M is a big number ($M = \sum_{u \in V} c(u)$). Let α be the maximum weight of a BIS of $\bar{G}_1$. Then the maximum weight of a BIS of G is $\alpha - r(W)M$.

4. Application to series-parallel graphs. These decomposition techniques are useful for classes of graphs that can be decomposed by two-node cutsets. For series-parallel graphs, Hassin and Tamir [12] proved the following result.

THEOREM 4.1. *Let $G = (V, E)$ be a series-parallel graph. There exist two nodes u, $v \in V$ and two subsets V_1, $V_2 \subseteq V$, such that*

(i) $|V_i| \leq \frac{2}{3}|V| + 2$, $i = 1, 2$,

(ii) $V_1 \cap V_2 = \{u, v\}$,

(iii) $V = V_1 \cup V_2$, $E = E(V_1) \cup E(V_2)$.

Furthermore, the vertices u, v and the sets V_1 and V_2 can be found in linear time.

To find a maximum weighted BIS, we should recursively decompose the graph. We might have to add five nodes to each piece so we do not decompose if the pieces are not smaller than the graph; i.e., we want $\frac{2}{3}|V| + 7 \leq |V|$. Thus we stop when each piece has at most 21 nodes. Figure 2 shows an example of a graph that cannot be further decomposed. If the graph has 21 nodes or less, we solve the problem by enumeration.

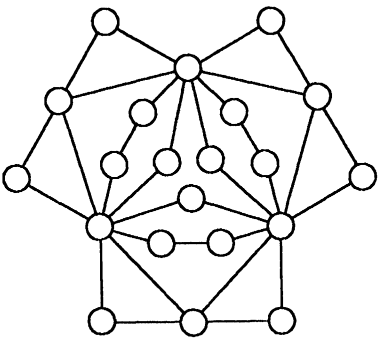


FIG. 2

Let $T(n)$ be the number of operations to solve the problem in a graph with n nodes. Then

$$T(n) \leq cn + T(n_1) + T(n_2),$$

where $n_i \leq \frac{2}{3}n + 2$, $i = 1, 2$ and $n_1 + n_2 = n + 2$. Therefore $T(n) = O(n \log n)$. We can state the following result.

THEOREM 4.2. *A maximum weighted* BIS *in a series-parallel graph with n nodes can be found in* $O(n \log n)$ *time.*

In [6] we showed that, even for series-parallel graphs, $P(G)$ may have facet-defining inequalities that are not simple to describe. However, the theorem below shows that, by allowing some extra variables, we obtain $P(G)$ as a projection of a polytope that is much easier to represent.

THEOREM 4.3. *If G is series-parallel and has n nodes, then* $P(G)$ *is a projection of a polytope defined by a system with* $O(n)$ *inequalities and* $O(n)$ *variables.*

Proof. Applying Theorems 2.1 and 2.2 to the decomposition of the graph, we obtain a polytope Q such that $P(G)$ is a projection of Q. If G has n nodes, then the number of variables in the system that defines Q is $O(n)$, and the number of inequalities is $O(n)$. For this, it is sufficient to know a characterization of $P(G)$ for series-parallel graphs with at most 21 nodes. □

5. Compositions of facets. Now we see that, in Cases 2 and 3 of §2, we obtain a complete description of the facets of $P(G)$ from the facets coming from the pieces. We must first study the structure of these inequalities.

Let $ax \leq \alpha$ be an inequality that defines a nontrivial facet of $P(G)$; i.e., a contains at least two nonzero components. It is easy to see that $a \geq 0$ and $\alpha > 0$. Also, if a has exactly two nonzero components, then it corresponds to $x(u) + x(v) \leq 1$. This can only be the case when u and v are linked by a positive edge and a negative edge. We denote by V_a the set

$$V_a = \{v \mid a_v > 0\}.$$

The graph $G_a = (V_a, E(V_a))$ is called a facet-inducing graph. We denote by $\beta(G)$ the set

$$\beta(G) = \{W \subseteq V \mid (W, E(W)) \text{ is balanced}\},$$

and β_a is the set

$$\beta_a = \{W \subseteq V \mid W \in \beta(G) \text{ and } ax^W = \alpha\}.$$

Given a path $u, u_1, u_2, \ldots, u_k, v$ between u and v, the nodes $u_1, \ldots, u_k$ are called *internal nodes.* Now we present several lemmas about the inequality $ax \leq \alpha$.

LEMMA 5.1. *The graph* G_a *is connected.*

Proof. Suppose that G_a is the union of two disjoint graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_1, E_2)$. Let a_1 (respectively, a_2) be the row obtained from a by setting to zero all the components associated with nodes in G_1 (respectively, G_2). Letting

$$\alpha_i = \max a_i x, \qquad x \in P(G), \quad i = 1, 2,$$

we have that $\alpha = \alpha_1 + \alpha_2$ and $a = a_1 + a_2$; thus $ax \leq \alpha$ can be obtained as the sum of two valid inequalities. This is not possible because it defines a facet. □

LEMMA 5.2. *If* G_a *contains a node u of degree* 2 *and its neighbours are v and w, then* $a_u \leq a_v$, $a_u \leq a_w$.

Proof. Since $ax \leq \alpha$ defines a nontrivial facet, there is a set $W \in \beta_a$ such that $u \notin W$. It implies that $v, w \in W$.

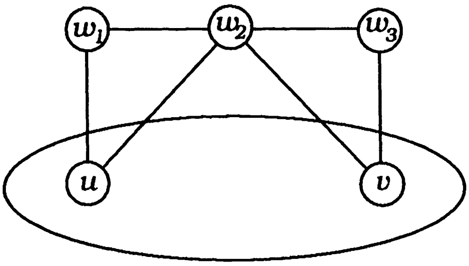


FIG. 3

Let

$$W' = W\backslash\{v\} \cup \{u\}, \qquad W'' = W\backslash\{w\} \cup \{u\}.$$

It is clear that $W' \in \beta(G)$ and $W'' \in \beta(G)$; then $a_u \leq a_v$, $a_u \leq a_w$. □

COROLLARY 5.3. *If G_a contains a path $u, u_1, \ldots, u_k, v$ whose internal nodes are of degree 2, then $a_{u_i} = a_{u_j}$ for $1 \leq i \leq k$, $1 \leq j \leq k$.*

LEMMA 5.4. *If G_a contains the induced subgraph $\Gamma = (U, E(U))$ where*

$$U = \{u, v, w_1, w_2, w_3\},$$

$$E(U) = \{uw_1, uw_2, vw_2, vw_3, w_1w_2, w_2w_3\},$$

and all the edges in $E(U)$ are labeled negative, then $a_{w_1} = a_{w_2} = a_{w_3}$. See Fig. 3.

Proof. Let $T_1 = \{u, w_1, w_2\}$ and $T_2 = \{v, w_2, w_3\}$. If $W \in \beta_a$, $W \cap \{u, w_2\} \neq \emptyset$, and $W \cap \{v, w_2\} \neq \emptyset$, then $|T_1 \cap W| = 2$ and $|T_2 \cap W| = 2$.

Since $ax \leq \alpha$ is different from the inequalities

$$x(u) + x(w_1) + x(w_2) \leq 2, \qquad x(v) + x(w_2) + x(w_3) \leq 2,$$

there are two node sets W_1 and W_2 in β_a such that $\{u, w_2\} \cap W_1 = \emptyset$, $\{v, w_2\} \cap W_2 = \emptyset$. Hence $w_1 \in W_1$ and $w_3 \in W_2$. Let

$$W'_1 = W_1\backslash\{w_1\} \cup \{w_2\}, \qquad W'_2 = W_2\backslash\{w_3\} \cup \{w_2\}.$$

Since $\{W'_1, W'_2\} \subseteq \beta(G)$, we have

$$a_{w_2} \leq a_{w_1}, \qquad a_{w_2} \leq a_{w_3}$$

and, from Lemma 5.2, we have

$$a_{w_2} \geq a_{w_1}, \qquad a_{w_2} \geq a_{w_3},$$

which yields

$$a_{w_1} = a_{w_2} = a_{w_3}. \quad □$$

LEMMA 5.5. *Given two nodes u and v, there is at most one path in G_a containing an even (respectively, odd) number of negative edges whose internal nodes are of degree 2 (a path could consist of a single edge).*

Proof. Suppose that G_a contains the paths $u, u_1, \ldots, u_k, v$ and $u, v_1, \ldots, v_l, v$ that satisfy the conditions above. If $W \in \beta_a$, then either (a) $|W \cap \{u_1, \ldots, u_k\}| = k$ and $|W \cap \{v_1, \ldots, v_l\}| = l$ or (b) $|W \cap \{u_1, \ldots, u_k\}| = k - 1$ and $|W \cap \{v_1, \ldots, v_l\}| = l - 1$. Thus

$$\sum_i x^W(u_i) - \sum_j x^W(v_j) = k - l,$$

but this is not possible because $ax \leq \alpha$ defines a facet. □

LEMMA 5.6. *Let p be a node of degree 3 in G_a; given any other node q in G_a, there is at most one path between p and q in G_a whose internal nodes are of degree 2.*

Proof. Suppose that there are two paths $p, u_1, \ldots, u_k, q$ and $p, v_1, \ldots, v_l, q$ that satisfy the above conditions. Because of Lemma 5.5, we assume that these paths have different parities.

Let $W \in \beta_a$. We have the two following cases:

(i) If $q \notin W$, then $\{p, u_1, \ldots, u_k, v_1, \ldots, v_l\} \subseteq W$, because p has degree 3 in G_a and $a \geq 0$;

(ii) If $q \in W$, then $|\{p, u_1, \ldots, u_k, v_1, \ldots, v_l\} \cap W| = k + l$, because $a \geq 0$ and W cannot contain a cycle with an odd number of negative edges.

Thus x^W satisfies

$$x(p) + x(q) + \sum_i x(u_i) + \sum_j x(v_j) = k + l + 1,$$

but this is not possible because $ax \leq \alpha$ defines a facet. □

LEMMA 5.7. *For a facet-defining inequality $ax \leq \alpha$, the graph G_a cannot be decomposed as in Theorem 2.1.*

Proof. Suppose that G_a admits such decomposition, since $ax \leq \alpha$ should also define a facet of $P(G_a)$. This contradicts Theorem 2.1. □

Now we study the facet-defining inequalities of $P(\bar{G}_k)$.

LEMMA 5.8. *If $ax \leq \alpha$ defines a facet of $P(\bar{G}_k)$, $k = 1, 2$ and $\{u, v, w_i;\ 1 \leq i \leq 4\} \subseteq V_a$ in Cases 2 and 3, $\{u, v, w_i;\ 1 \leq i \leq 5\} \subseteq V_a$, in Case 4, then $ax \leq \alpha$ is of the type (2.1) or (2.2).*

Proof. In Case 2, we can apply Lemma 5.4 and we have that $a_{w_2} = a_{w_3} = a_{w_4}$.

Let $W \in \beta_a$. There are the three following cases:

(i) $W \cap \{u, v\} = \emptyset$; in this case, $\{w_i \mid 1 \leq i \leq 4\} \subseteq W$;

(ii) $\{u, v\} \subseteq W$; in this case, $w_1 \notin W$, because it would create a cycle with one negative edge, and $W \cap \{w_2, w_3, w_4\} = \{w_2, w_4\}$, because $a_{w_2} = a_{w_3} = a_{w_4}$;

(iii) $|\{u, v\} \cap W| = 1$; this implies that $|W \cap \{w_2, w_3, w_4\}| = 2$; also, $w_1 \in W$, because $a_{w_1} > 0$.

Therefore $|W \cap \{u, v, w_i;\ 1 \leq i \leq 4\}| = 4$. Hence x^W satisfies (2.1); this implies that $ax \leq \alpha$ is of the type (2.1). In Case 3 or Case 4, the proof is analogous. □

Consider now a nontrivial facet-defining inequality $ax \leq \alpha$. In Cases 2 and 3, the structure of G_a falls into one of the types below:

(i) G_a does not contain any of $\{w_i \mid 1 \leq i \leq 4\}$;

(ii) If $w_2 \in G_a$, Lemma 5.7 shows that $u, w_3 \in G_a$; if there is any other node in G_a, then Lemma 5.7 shows that $v \in G_a$, and Lemma 5.6 shows that w_3 should not have degree 3. Hence $w_4 \in G_a$. From Lemma 5.4, we have that $a_{w_2} = a_{w_3} = a_{w_4}$. We can assume without loss of generality that $a_{w_2} = 1$. From Lemma 5.8, we have that $w_1 \in G_a$ only if the inequality is of type (2.1).

(iii) If $\{w_2, w_4\} \cap V_a = \emptyset$, then Lemma 5.5 shows that, in Case 2, w_3 cannot be in V_a; however, w_1 could be in V_a. In Case 3, Lemma 5.5 shows that $|\{w_1, w_3\} \cap V_a| \leq 1$.

Thus we have that, in Cases 2 and 3, the facet-defining inequalities of $P(\bar{G}_k)$ are classified as follows, for $k = 1, 2$:

$$\sum_{j \in V_k} a_{ij}^k x(j) \leq \alpha_i^k, \qquad i \in I_1^k, \tag{5.1a}$$

$$\sum_{j \in V_k} a_{ij}^k x(j) + x(w_2) + x(w_3) + x(w_4) \leq \alpha_i^k, \qquad i \in I_2^k, \tag{5.1b}$$

$$\sum_{j \in V_k} a_{ij}^k x(j) + x(w_3) \le \alpha_i^k, \qquad i \in I_3^k, \tag{5.1c}$$

$$\sum_{j \in V_k} a_{ij}^k x(j) + x(w_1) \le \alpha_i^k, \qquad i \in I_4^k, \tag{5.1d}$$

$$x(u) + x(w_2) + x(w_3) \le 2, \tag{5.1e}$$

$$x(v) + x(w_3) + x(w_4) \le 2, \tag{5.1f}$$

$$x(u) + x(v) + x(w_1) + x(w_2) + x(w_3) + x(w_4) \le 4, \tag{5.1g}$$

$$x(w_j) \le 1, \qquad 1 \le j \le 4, \tag{5.1h}$$

$$x(j) \ge 0, \qquad j \in \bar{V}_k. \tag{5.1i}$$

Then $F(\bar{G})$ is defined by both systems, together with the inequality

$$-x(u) - x(v) - x(w_1) - x(w_2) - x(w_3) - x(w_4) \le -4. \tag{5.2}$$

To project the variables $\{x(w_i)\}$, we use the following theorem of Balas and Pulleyblank [1].

THEOREM 5.9. *Let $Z = \{(w, x) \mid Aw + Bx \le b\}$; the projection of Z along the subspace of the w variables is*

$$X = \{x \mid (vB)x \le vb,\ \forall v \in \text{extr } Y,\ x \ge 0\},$$

where extr Y *denotes the set of extreme rays of*

$$Y = \{y \mid yA \ge 0,\ y \ge 0\}.$$

In our case, the matrix A has the following twelve types of rows:

$$\begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
-1 & -1 & -1 & -1\,.
\end{array}$$

Column j corresponds to $x(w_j)$ for $1 \le j \le 4$; the first seven rows correspond to inequalities (5.1a)–(5.1g), respectively; the next four rows correspond to inequalities (5.1h); the last row corresponds to inequality (5.2).

The extreme rays of Y correspond to the extreme points of

$$\{y \mid yA \ge 0,\ \sum y_i = 1,\ y \ge 0\},$$

so we enumerate the extreme points of

$$\{z \mid Bz \ge 0,\ \sum z_i = 1,\ z \ge 0\},$$

where B is the matrix below:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & -1 \end{bmatrix}.$$

The extreme points are the columns of

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & & \vdots & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ & 0 & \ddots & & 0 & 0 & 0 & 0 & 0 & \frac{1}{5} \\ & & & & \frac{1}{3} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{5} \\ \vdots & \vdots & & & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ & & & & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\ & & & & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ & & & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{5} \\ & & & 1 & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{5} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{5} \end{bmatrix}.$$

So the inequalities given by Theorem 5.9 are obtained by performing the following steps:

(i) Keep inequalities (5.1a)–(5.1g), but delete the variables $\{x(w_i)\}$;

(ii) Add three inequalities, one of type (5.1b), one of type (5.1d) or $x(w_1) \leq 1$, and (5.2),

(iii) Add four inequalities, one of type (5.1d) or $x(w_1) \leq 1$, one of type (5.1e), one of type (5.1f), and (5.2), from the result delete $x(w_3)$;

(iv) Add four inequalities, one of type (5.1d) or $x(w_1) \leq 1$, one of type (5.1e), $x(w_4) \leq 1$, and (5.2);

(v) Add four inequalities, one of type (5.1d) or $x(w_1) \leq 1$, one of type (5.1f), $x(w_2) \leq 1$, and (5.2);

(vi) Add (5.1g) and (5.2) (which gives the redundant inequality $0 \leq 0$);

(vii) Add five inequalities, one of type (5.1c) or $x(w_3) \leq 1$, one of type (5.1d) or $x(w_1) \leq 1$, $x(w_2) \leq 1$, $x(w_4) \leq 1$, and (5.2).

The next lemma shows that some of these inequalities are redundant.

LEMMA 5.10. *Let $ax \leq \alpha$ be a facet-defining inequality of $P(G)$, $\alpha \geq 0$. If $V_a \subseteq V_k$, then this inequality also defines a facet of $P(\bar{G}_k)$, $k = 1, 2$.*

Proof. It is clear that this inequality is valid for $P(\bar{G}_k)$.

By hypothesis, there are $|\bar{V}_k|$ linearly independent incidence vectors of balanced induced subgraphs that satisfy $ax = \alpha$. Let us form a matrix M with them. Among these vectors, there is one vector $\bar{x}$ with $\bar{x}(u) = 0$ or $\bar{x}(v) = 0$. Hence the columns of

$$M' = \begin{bmatrix} M & \bar{x} & \bar{x} & \bar{x} & \bar{x} \\ 0 & 1 & & & \\ 0 & & 1 & & \\ 0 & & & 1 & \\ 0 & & & & 1 \end{bmatrix}$$

form a set of $|\bar{V}_k|$ linearly independent incidence vectors of balanced induced subgraphs of $\bar{G}_k$. Moreover, they satisfy $ax = \alpha$. □

This lemma states that, if an inequality is essential in the definition of $P(G)$ and its support is included in V_k, then a multiple of this inequality already appears in (5.1a). So in (i) we should only keep (5.1a) and in (ii) we should keep the constraints whose support intersects both $V_1 \backslash \{u, v\}$ and $V_2 \backslash \{u, v\}$. For the same reasons, the inequalities produced in (iii)–(v) are redundant. Thus we obtain the following result.

THEOREM 5.11. *In Cases* 2 *and* 3, *$P(G)$ is defined by* (5.1a), *together with $x(j) \geq 0$ for $j \in V$ and the mixed inequalities*

$$(5.3) \qquad \sum_{j \in V_k} a_{ij}^k x(j) + \sum_{j \in V_l} a_{sj}^l x(j) - x(u) - x(v) \leq \alpha_i^k + \alpha_s^l - 4$$

for $k = 1, 2$; $l = 1, 2$; $i \in I_2^k$, $s \in I_4^l$.

In what follows, we prove that, if $k \neq l$, inequalities (5.3) define facets of $P(G)$. First, we must introduce a technical lemma.

LEMMA 5.12. *Inequalities* (5.1b) *and* (5.1d) *define facets of $F(\bar{G}_k)$.*

Proof. First, let us study inequalities (5.1b). Since $\dim (F(\bar{G}_k)) = |\bar{V}_k| - 1$, we must show that there are $|\bar{V}_k| - 1$ linearly independent vectors in $F(\bar{G})$ that satisfy (5.1b). Let $n = |\bar{V}_k|$. There is a linearly independent set $\{x_1, \ldots, x_n\}$ of extreme points of $P(\bar{G}_k)$ that satisfy (5.1b).

For a vector x_j, we do the following. If x_j satisfies (5.2), we keep it; otherwise, we set the component $x_j(w_1)$ equal to 1. We obtain a set $S = \{x'_1, \ldots, x'_n\}$ of vectors that satisfy both (5.1b) and (5.2). Since we only modified one component, there are $n - 1$ linearly independent vectors in S.

Now let us study an inequality of type (5.1d), say $ax \leq \alpha$. Let $S = \{x_1, \ldots, x_n\}$ be a set of linearly independent extreme points of $P(\bar{G}_k)$ that satisfy $ax = \alpha$. Let x_j be one of these vectors; if x_j satisfies (5.2), we define $x'_j = x_j$; otherwise, we define $x'_j(u) = x_j(u)$ if $u \neq w_2, w_3, w_4$. We set to 1 some of the components $x'_j(w_2)$, $x'_j(w_3)$, $x'_j(w_4)$ to obtain an extreme point x'_j that satisfies (5.2).

Let us assume that $ax \leq \alpha$ is inessential in the definition of $F(\bar{G})$. This implies that

$$a = \sum \lambda_i b_i + \gamma d, \quad \alpha \geq \sum \lambda_i \beta_i + \gamma \varepsilon, \quad \lambda_i \geq 0,$$

where $b_i x \leq \beta_i$ denotes an inequality in (5.1), other than $ax \leq \alpha$, but not (5.1g), and $dx = \varepsilon$ denotes the equation derived from (5.1g).

For each inequality $b_i x \leq \beta_i$, there is a vector $x_j \in S$ such that $b_i x_j < \beta_i$ and $b_i x'_j = \beta_i$. Hence $b_{iw_j} > 0$, $j = 2, 3, 4$; then $b_{iw_1} = 0$. This implies $\gamma = 1$ and some $\lambda_i < 0$, a contradiction. □

So if $ax \leq \alpha$ is of type (5.1b) or (5.1d) and another inequality $bx \leq \beta$ defines the same face of $F(\bar{G}_k)$, then $b = \lambda a + \gamma d$, where $dx = \varepsilon$ denotes (5.1g) as equation and $\lambda \geq 0$. The constraint $bx \leq \beta$ cannot be in (5.1), because of the structure of the inequalities in this system.

LEMMA 5.13. *If $k \neq l$, inequalities* (5.3) *define facets of $P(G)$.*

Proof. Suppose that $k = 1$, $l = 2$. We show that there exists a vector $\bar{x}$ in $P(G)$ that satisfies this inequality as equation, and all the others as strict inequality.

For the inequality

$$\sum_{j \in V_1} a_{ij}^1 x(j) + x(w_2) + x(w_3) + x(w_4) \leq \alpha_i^1, \qquad i \in I_2^1,$$

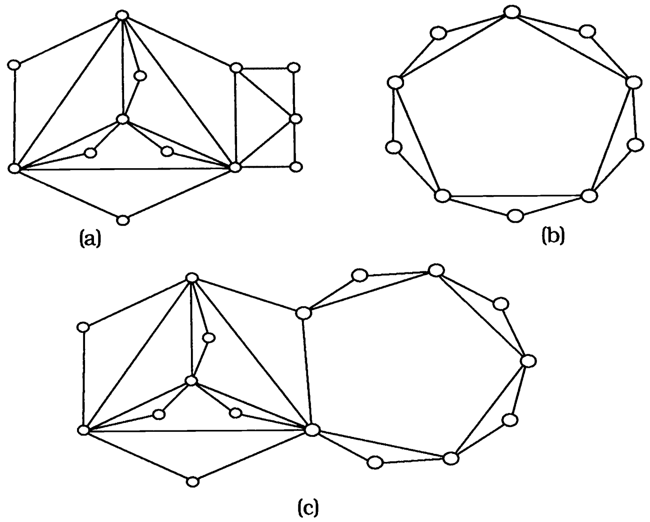


FIG. 4

let $S = \{x_1, \dots, x_l\}$ be the set of extreme points of $F(\bar{G}_1)$ that satisfy it as equation. For

$$\sum_{j\in V_2} a_{ij}^2 x(j) + x(w_1) \le \alpha_i^2, \qquad i \in I_4^2,$$

let $T = \{y_1, \dots, y_m\}$ be the set of extreme points of $F(\bar{G}_2)$ that satisfy it as equation.

First, for each vector x_i, we can find a vector y_{j_i} such that both together give a vector in $F(\bar{G})$, $0 \le i \le l$. Next, for each vector y_j, we can find a vector x_{i_j} such that together they give a vector in $F(\bar{G})$, $0 \le j \le m$.

Let $\{z_1, \dots, z_r\}$ be the set of vectors thus obtained ($r = l + m$). Let z_k' be the vector obtained by dropping the components $z_k(w_j)$, $1 \le j \le 4$.

Then

$$\bar{x} = \frac{1}{r}(z_1' + \cdots + z_r')$$

is the required vector. □

This theorem gives a way to describe facets of $P(G)$ by composition of facets for the pieces. For instance, consider the graphs in Figs. 4(a) and 4(b), with all the edges labeled negative. The inequalities

$$\sum x(i) \le 9 \quad \text{and} \quad \sum x(i) \le 7$$

define facets for the first and second graph respectively; see [6]. Theorem 5.16 shows that

$$\sum x(i) \le 12$$

defines a facet for the graph in Fig. 4(c).

The techniques of this section also apply to Case 4 of §2. The only missing piece in this case is a characterization of the extreme rays of the set Y, defined in Theorem 5.9.

6. Acyclic induced subgraphs. In this section, our aim is to show how the same ideas apply to the acyclic induced subgraph polytope. Similar compositions for the polytope of acyclic spanning subgraphs have been studied in [5].

Let $D = (V, A)$ be a directed graph; the induced subgraph $(W, A(W))$ is called acyclic if it does not have a directed cycle. The *acyclic induced subgraph* (AIS) polytope is

$$P'(D) = \text{conv}\ \{x^W \in \Re^V \mid (W, A(W)) \text{ is acyclic}\},$$

and the *maximum* AIS *problem* is

$$\max cx, \quad x \in P'(D).$$

If $G = (V, E)$ is an undirected graph, the maximum stable set problem in G can be reduced to a maximum AIS problem in a directed graph $D = (V, A)$, where each edge $e \in E$ is replaced by the arcs (i, j) and (j, i). This shows that the maximum AIS problem is NP-hard for planar digraphs.

The analogue of Theorem 2.1 is the following result.

THEOREM 6.1. *Let $D = (V, A)$ be a directed graph such that there exist two node sets V_1 and V_2 with the following properties*:

(i) $V = V_1 \cup V_2$,

(ii) $W = V_1 \cap V_2 \neq \emptyset$,

(iii) *For $\{i, j\} \subseteq W$, the arc $(i, j) \in A$ and $(j, i) \in A$*,

(iv) *The induced subgraph $(V \backslash W, A(V \backslash W))$ is disconnected.*

If $D_1 = (V_1, A(V_1))$ and $D_2 = (V_2, A(V_2))$, then a system of inequalities that defines $P'(D)$ is obtained by the juxtaposition of such systems defining $P'(D_1)$ and $P'(D_2)$.

Now let us study digraphs with a two-vertex cutset.

Let $D_1 = (V_1, A_1)$ and $D_2 = (V_2, A_2)$ be two digraphs such that $V_1 \cap V_2 = \{u, v\}$ and let $D = (V, A)$ be the union of D_1 and D_2, i.e., $V = V_1 \cup V_2$, $A = A_1 \cup A_2$. There are three cases.

Case 1. The arcs (u, v) and (v, u) belong to A.

Case 2. The arc $(u, v) \in A$.

Case 3. There is no arc between u and v.

Case 1 is covered by Theorem 6.1. In Case 2, we define $\bar{D}_i = (\bar{V}_i, \bar{A}_i)$, $i = 1, 2$ as follows:

$$\bar{V}_i = V_i \cup \{w_1, w_2\},$$

$$\bar{A}_i = A_i \cup \{(w_1, u), (v, w_1), (u, w_2), (w_2, u), (v, w_2), (w_2, v)\}.$$

In Case 3, we define

$$\bar{V}_i = V_i \cup \{w_1, w_2, w_3\},$$

$$\bar{A}_i = A_i \cup \{(v, w_1), (w_1, u), (u, w_2), (w_2, v), (u, w_3), (w_3, u), (w_3, v), (v, w_3)\}.$$

For Case 2, the inequality

$$x(w_1) + x(w_2) + x(u) + x(v) \leq 2$$

plays the role of inequality (2.1). For Case 3, the inequality

$$x(w_1) + x(w_2) + x(w_3) + x(u) + x(v) \leq 3$$

plays the role of (2.2).

So they define a facet $F(\bar{D}_i)$ of $P'(\bar{D}_i)$, $i = 1, 2$ and a facet $F(\bar{D})$ of $P'(\bar{D})$; again, the polytope $P'(D)$ is the projection of $F(\bar{D})$ along the variables $\{x(w_i)\}$, and the following is the analogue of Theorem 2.2.

THEOREM 6.2. *The juxtaposition of a system that defines $F(\bar{D}_1)$ and a system that defines $F(\bar{D}_2)$ gives a system that defines $F(\bar{D})$.*

Algorithmic aspects analogous to those of §3 hold for the AIS problem. For series-parallel digraphs, we have the following result.

THEOREM 6.3. *If D is a series-parallel directed graph with n nodes, then a maximum weighted* AIS *can be found in $O(n \log n)$ time.*

THEOREM 6.4. *If D is a series-parallel directed graph with n nodes, then $P'(D)$ is a projection of a polytope defined by a system with $O(n)$ inequalities and $O(n)$ variables.*

As for the BIS polytope, $P'(D)$ may have facet-defining inequalities that are not easy to describe even for series-parallel digraphs [6]. The above theorem shows that, if we allow extra variables, then we have a polytope that has a much simpler representation.

The techniques of §5 can also be adapted to Case 2 of the present section to produce compositions of facets of $P(D')$.

Acknowledgment. We are grateful to the referee for his suggestions on the presentation of this paper.

REFERENCES

[1] E. BALAS AND W. R. PULLEYBLANK, *The perfectly matchable subgraph polytope of a bipartite graph*, Networks, 13 (1983), pp. 495–516.

[2] F. BARAHONA, *Balancing Signed Toroidal Graphs in Polynomial Time*, Depto. de Matemáticas, Universidad de Chile, 1981.

[3] ———, *The max cut problem in graphs not contractible to K_5*, Oper. Res. Lett., 2 (1983), pp. 107–111.

[4] F. BARAHONA AND A. R. MAHJOUB, *On the cut polytope*, Math. Programming, 36 (1986), pp. 157–173.

[5] F. BARAHONA, J. FONLUPT, AND A. R. MAHJOUB, *Compositions of graphs and polyhedra* IV: *Acyclic spanning subgraphs*, SIAM J. Discrete Math.. 7 (1994), pp. 390–402. this issue.

[6] F. BARAHONA AND A. R. MAHJOUB, *Facets of the balanced (acyclic) induced subgraph polytope*, Math. Programming, 45 (1989), pp. 21–33.

[7] V. CHVÁTAL, *On certain polytopes associated with graphs*, J. Combin. Theory Ser. B, 18 (1975), pp. 138–154.

[8] G. CORNUÉJOLS, D. NADDEF, AND W. R. PULLEYBLANK, *The traveling salesman problem in graphs with 3-edge cutsets*, J. Assoc. Comput. Mach., 32 (1985), pp. 383–410.

[9] R. EULER AND A. R. MAHJOUB, *On a composition of independence systems by circuit identification*, J. Combin. Theory Ser. B, 53 (1991), pp. 235–259.

[10] J. FONLUPT, A. R. MAHJOUB, AND J. P. UHRY, *Compositions in the bipartite subgraph polytope*, Discrete Math., to appear.

[11] F. HARARY, *On the notion of balance of a signed graph*, Mich. Math. J., 2 (1952), pp. 143–146.

[12] R. HASSIN AND A. TAMIR, *Efficient algorithms for optimization and selection on series-parallel graphs*, SIAM J. Algebraic Discrete Math., 7 (1986), pp. 379–389.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 359–371, August 1994

003

COMPOSITIONS OF GRAPHS AND POLYHEDRA II: STABLE SETS*

FRANCISCO BARAHONA[†] AND ALI RIDHA MAHJOUB[‡]

Abstract. A graph G with a two-node cutset decomposes into two pieces. A technique to describe the stable set polytope for G based on stable set polytopes associated with the pieces is studied. This gives a way to characterize this polytope for classes of graphs that can be recursively decomposed. This also gives a procedure to describe new facets of this polytope. A compact system for the stable set problem in series-parallel graphs is derived. This technique is also applied to characterize facet-defining inequalities for graphs with no $K_5\backslash e$ minor. The stable set problem is polynomially solvable for this class of graphs. Compositions of h-perfect graphs are also studied.

Key words. polyhedral combinatorics, composition of polyhedra, stable set polytope, compact systems

AMS subject classifications. 05C85, 90C27

1. Introduction. Given a graph G, let $P(G)$ be the stable set polytope of G. If G has a one- or two-node cutset, then G decomposes into G_1 and G_2. We study a technique to derive a system of inequalities that defines $P(G)$ from systems related to G_1 and G_2. In a companion paper [2], we studied the same technique for the polytopes of balanced and acyclic subgraphs. We can use this to characterize the stable set polytope for classes of graphs that can be decomposed by two-vertex cuts, provided that the pieces are "easy" to handle. It also gives a procedure for characterizing facets of the stable set polytope by composition of facets for the pieces. We use this method in [3] to characterize the stable set polytope for graphs with no W_4 minor.

In §2 we study the structure of the facets of $P(G)$ and show some facet-defining inequalities for subdivisions of a wheel. In §3 we study the composition of polyhedra. In §4 we study the algorithmic aspects of this kind of composition. In §5 we study series-parallel graphs. We derive a compact system for the stable set problem in this class of graphs; i.e., we show that $P(G)$ is a projection of a polyhedron that is defined by a system whose number of variables and number of inequalities is linear in the number of nodes of the graph. In §6 we study some facets of $P(G)$ for graphs with no $K_5\backslash e$ minor. Based on a decomposition theorem of Wagner, we can derive a polynomial algorithm for finding a maximum weighted stable set in this class of graphs. Using composition of facets, we show that, for any positive integer p, we can find a graph G with no $K_5\backslash e$ minor such that $P(G)$ has a facet-defining inequality with coefficients $1, 2, \ldots, p$. In §7 we study compositions of h-perfect graphs.

We finish this introduction with a few definitions. Given a graph $G = (V, E)$, a stable set $S \subseteq V$ is a node set such that there is no edge with both endnodes in S. If $S \subseteq V$, let $x^S \in \Re^V$, where $x^S(u) = 1$ if $u \in S$, and $x^S(u) = 0$ if $u \notin S$; x^S is called the *incidence vector* of S.

* Received by the editors June 4, 1990; accepted for publication (in revised form) May 20, 1990. Part of this work was done at the Institute for Operations Research of the University of Bonn and supported by the joint project "Combinatorial Optimization" of the Natural Sciences and Engineering Research Council of Canada and the German Research Association (Deutsche Forschungsgemeinschaft, SFB 303).

[†] IBM T. J. Watson Research Center, Yorktown Heights, New York 10598.

[‡] Department of Statistics, King Saud University, Riyadh, Saudi Arabia. Present address, Laboratoire d'Informatique de Brest (LIBr), Université de Bretagne Occidentale, 6 Avenue Le Gorgeu, 29287 Brest Cedex, France. This author was supported by C. P. Rail.

The *stable set polytope* $P(G)$ is the convex hull of incidence vectors of all stable sets of G, i.e.,

$$P(G) = \text{conv}\,\{x^S \in \Re^V \,|\, S \text{ is a stable set of } G\}.$$

The polytope $P(G)$ is full-dimensional. This implies that (up to multiplication by a positive constant) there is a unique nonredundant inequality system $Ax \leq b$ such that $P(G) = \{x \,|\, Ax \leq b\}$; moreover, there is a natural bijection among the facets of $P(G)$ and the inequalities of that system.

2. On the facets of $P(G)$. The facets of $P(G)$ have been studied in [11], [5], [10], [13]–[15]. In this section, we present some properties of those inequalities that will be used later. We also present some facets for subdivisions of a wheel.

Let $ax \leq \alpha$ be an inequality that defines a facet of $P(G)$. If a contains at least two nonzero components, we say that $ax \leq \alpha$ defines a nontrivial facet. In this section, we study only nontrivial facets, so we have that $a \geq 0$ and $\alpha > 0$. We denote by V_a the set

$$V_a = \{v \,|\, a_v > 0\}.$$

The subgraph induced by V_a is denoted by G_a. Let us remark that G_a is a two-connected graph.

We now present two lemmas about the structure of G_a; their proofs appear in [9].

LEMMA 2.1. *If G_a contains a path with vertices p, u, v, q, where u and v are of degree* 2, *then $a_u = a_v$.*

LEMMA 2.2. *If G_a is different from an odd hole* (*and from K_3*), *then it does not contain between two given nodes p and q two edge-disjoint paths such that each node of them different from p, q is of degree* 2.

In what follows, we give two procedures of construction of facets of the stable set polytope from known facets. The first procedure consists of subdividing a star.

THEOREM 2.3 (subdivision of a star). *Let $G = (V, E)$ be a graph and $ax \leq \alpha$ be a nontrivial facet-defining inequality. Let v be a vertex of G and $N = \{v_0, \ldots, v_{k-1}\}$ be the neighbor set of v. Suppose that, for each $i = 0, \ldots, k-1$, there exists a stable set $\tilde{S}_i$ such that $ax^{\tilde{S}_i} = \alpha$ and $\tilde{S}_i \cap N = \{v_i, v_{i+1}, \ldots, v_{i+p-1}\}$, where $p \geq 1$ is a fixed integer and the indices are numbered modulo k. Suppose also that p and k are relatively prime and $a_{v_0} = a_{v_1} = \cdots = a_{v_{k-1}} = a_v/p$. Let $G' = (V', E')$ be the graph obtained from G by adding on each edge vv_i a new node v_i' for $i = 0, \ldots, k-1$. Set*

$$\bar{a}_u = a_u \quad \text{for } u \in V\backslash\{v\},$$

$$\bar{a}_v = a_v(k-p)/p,$$

$$\bar{a}_{v_i'} = a_v/p \quad \text{for } i = 0, 1, \ldots, k-1,$$

$$\bar{\alpha} = \alpha + a_v(k-p)/p.$$

Then $\bar{a}x \leq \bar{\alpha}$ defines a facet of $P(G')$.

Proof. First, let us show that $\bar{a}x \leq \bar{\alpha}$ is valid for $P(G')$. Let S' be a maximal stable set of G'.

Case 1. The node v belongs to S'.

Then $S = S'\backslash\{v\}$ is a stable set in G, which implies that $ax^S \leq \alpha$ and then $\bar{a}x^{S'} \leq \bar{\alpha}$.

Case 2. The node v does not belong to S'.

Let $T = S' \cap N$ and $T' = \{v_i' \,|\, v_i \notin T,\ 0 \leq i \leq k-1\}$. Note that $T' \subseteq S'$. Let $S = (S'\backslash(T \cup T')) \cup \{v\}$. It is clear that S is a stable set of G; then $ax^S \leq \alpha$. Since $|T \cup T'| = k$, we have that $\bar{a}x^{S'} \leq \bar{\alpha}$.

Now we have to show that $\bar{a}x \leq \bar{\alpha}$ is facet-inducing. Let $n = |V|$ and $m = n + k$. There are n stable sets $S_1, \ldots, S_n$ of G whose incidence vectors are linearly independent and satisfy $ax = \alpha$. Consider the following sets:

$$S_i' = (S_i \backslash \{v\}) \cup \{v_0', \ldots, v_{k-1}'\} \quad \text{if } v \in S_i,$$
$$S_i' = S_i \cup \{v\} \quad \text{if } v \notin S_i, \text{ for } i = 1, \ldots, n,$$

and

$$S_{n+j+1}' = \tilde{S}_j \cup \{v_i' | v_i \notin \tilde{S}_j\} \quad \text{for } j = 0, 1, \ldots, k-1.$$

The incidence vectors of $S_1', \ldots, S_m'$ satisfy $\bar{a}x = \bar{\alpha}$. Let us assume that they also satisfy $bx = \bar{\alpha}$, where $bx \leq \bar{\alpha}$ is a facet-defining inequality of $P(G')$. We prove that $b = \bar{a}$.

Since the incidence vectors of $\tilde{S}_0, \ldots, \tilde{S}_{k-1}$ are linearly independent, we can assume that $S_1 = \tilde{S}_0, \ldots, S_k = \tilde{S}_{k-1}$.

Consider the equations $bx^{S_i'} - bx^{S_{n+i}'} = 0$, $i = 1, \ldots, k$. This is a system like

$$[\lambda \gamma] \begin{pmatrix} u \\ -C \end{pmatrix} = 0,$$

where u is a row of 1's, and C is the $k \times k$ cyclic matrix having $(k - p)$ 1's in each row and column. Hence, we have that

$$b_{v_i'} = b_v/(k-p) \quad \text{for } i = 0, \ldots, k-1.$$

There is some number $\delta > 0$ such that $\bar{\alpha} - b_v = \delta\alpha$.

Consider the equations $bx^{S_i'} = \bar{\alpha}$ (or $bx^{S_i'} - b_v = \bar{\alpha} - b_v$), $i = 1, \ldots, n$. Since a is the unique solution of $ax^{S_i} = \alpha$, $i = 1, \ldots, n$, and $\bar{\alpha} - b_v = \delta\alpha$, we have that

$$b_u = \delta a_u \quad \text{for } u \in V \backslash \{v\} \quad \text{and} \quad b_v p/(k-p) = \delta a_v.$$

Therefore,

$$b_v = \delta a_v (k-p)/p = \delta \bar{a}_v \quad \text{and} \quad b_{v_i'} = \delta a_v/p = \delta \bar{a}_{v_i'}, \quad i = 0, \ldots, k-1.$$

Since $\bar{\alpha} = \delta\alpha + b_v = \delta\alpha + \delta a_v(k-p)/p$, we have that $\delta = 1$. The proof is complete. □

Wolsey [15] gave some methods to construct facets of $P(G)$ from known ones. One of those methods is the following, which consists of replacing one edge by a chordless path of length 3.

THEOREM 2.4 (subdivision of an edge). *Given a graph $G = (V, E)$ and $uv \in E$, let $ax \leq \alpha$ be a nontrivial facet-defining inequality of $P(G)$, different from $x(u) + x(v) \leq 1$. Let G' be the graph G without the edge uv, if $\beta = \max\{ax | x \in P(G')\}$ has a solution with $x(u) = x(v) = 1$, then*

$$ax + \lambda x(s) + \lambda x(t) \leq \beta$$

defines a facet of $P(G'')$, where $\lambda = \beta - \alpha$, and G'' has been obtained by adding the nodes s and t to G', and the edges us, st, and tv.

In the following, we show a converse transformation.

THEOREM 2.5. *Let $G = (V, E)$ be a graph. Let $ax \leq \alpha$ be a facet-defining inequality of $P(G)$. Suppose that G contains a path (pu, uv, vq) such that u and v are of degree 2. Assume also that $a_p = a_u = a_v = \beta$. Let $G' = (V', E')$ be the graph obtained from G by* *replacing that path by the edge pq. Let*

$$\bar{a}_u = a_u \quad \textit{for } u \in V',$$
$$\bar{\alpha} = \alpha - \beta,$$

then $\bar{a}x \le \bar{\alpha}$ *defines a face of* $P(G')$.

Proof. First, we show that $\bar{a}x \le \bar{\alpha}$ is valid for $P(G')$. Let S' be a stable set of G'. If $\{p, q\} \cap S' \neq \emptyset$, say $p \in S'$, then $S = S' \cup \{v\}$ is a stable set in G; hence $\bar{a}x^{S'} \le \alpha - \beta = \bar{\alpha}$. If $\{p, q\} \cap S' = \emptyset$, then $S = S' \cup \{v\}$ is a stable set of G and thus $\bar{a}x^{S'} \le \alpha - \beta = \bar{\alpha}$.

Let $n = |V|$ and $m = n - 2$. We must exhibit m stable sets of G' whose incidence vectors are linearly independent and satisfy $\bar{a}x = \bar{\alpha}$.

Since $ax \le \alpha$ defines a facet of $P(G)$, there are n stable sets $S_1, \dots, S_n$ of G such that $ax^{S_i} = \alpha$, $1 \le i \le n$, and this set of vectors is linearly independent. Consider the following sets:

1) $S_i' = S_i \backslash \{v\}$ if $\{p, v\} \subseteq S_i$,
2) $S_i' = S_i \backslash \{p\}$ if $\{p, q\} \subseteq S_i$,
3) $S_i' = S_i \backslash \{u\}$ if $\{q, u\} \subseteq S_i$,
4) $S_i' = S_i \backslash \{u\}$ if $u \in S_i$, $q \notin S_i$,
5) $S_i' = S_i \backslash \{v\}$ if $v \in S_i$, $p \notin S_i$

for $i = 1, \dots, n$.

Note that the sets S_i' for $i = 1, \dots, n$ are all stable sets of G'. Let us denote by M (respectively, M') the matrix whose columns are the incidence vectors of $S_1, \dots, S_n$ $(S_1', \dots, S_m')$. The matrices M and M' look like

$$M = \left(\begin{array}{ccccc} A_1 & A_2 & A_3 & A_4 & A_5 \\ 1\cdots1 & 1\cdots1 & 0\cdots0 & 0\cdots0 & 0\cdots0 \\ 0\cdots0 & 0\cdots0 & 1\cdots1 & 1\cdots1 & 0\cdots0 \\ 1\cdots1 & 0\cdots0 & 0\cdots0 & 0\cdots0 & 1\cdots1 \\ 0\cdots0 & 1\cdots1 & 1\cdots1 & 0\cdots0 & 0\cdots0 \end{array}\right),$$

$$M' = \left(\begin{array}{ccccc} A_1 & A_2 & A_3 & A_4 & A_5 \\ 1\cdots1 & 0\cdots0 & 0\cdots0 & 0\cdots0 & 0\cdots0 \\ 0\cdots0 & 1\cdots1 & 1\cdots1 & 0\cdots0 & 0\cdots0 \end{array}\right).$$

We must show that the rank of M' is $n - 2$. Let $\bar{M}$ be the following matrix:

$$\bar{M} = \begin{pmatrix} & 0 \\ & \vdots \\ M & 0 \\ & 1 \\ & 0 \\ & 0 \\ & 0 \\ 1\cdots & 1 \end{pmatrix}.$$

This is nonsingular. In fact, if $\bar{M}$ is singular, then its last row should be linearly dependent of the others. Since a is the only solution of $tM = (\alpha, \dots, \alpha)$, we should have $\beta/\alpha = 1$. However, $\alpha \ge 2\beta$, a contradiction.

Now let us add the rows corresponding to u and v to the row corresponding to p and subtract from the resulting row the last row of $\bar{M}$. We obtain the following:

$$\left(\begin{array}{cccccc} A_1 & A_2 & A_3 & A_4 & A_5 & 0 \\ 1\cdots1 & 0\cdots0 & 0\cdots0 & 0\cdots0 & 0\cdots0 & \vdots \\ 0\cdots0 & 0\cdots0 & 1\cdots1 & 1\cdots1 & 0\cdots0 & 0 \\ 1\cdots1 & 0\cdots0 & 0\cdots0 & 0\cdots0 & 1\cdots1 & 0 \\ 0\cdots0 & 1\cdots1 & 1\cdots1 & 0\cdots0 & 0\cdots0 & 0 \\ 1111\ \cdots & & & & \cdots & 1 \end{array}\right).$$

Since this matrix is nonsingular, we can conclude that M' is of rank $n-2$. □

We finish this section by showing some facet-defining inequalities of $P(G)$, when G is a subdivision of a wheel.

Let G be the graph of Fig. 1(a); it is well known that the inequality

$$\sum_{j=1}^{5} x(j) + 2x(6) \leq 2$$

defines a facet of $P(G)$. By applying Theorem 2.3 to the star of node 6 and then to the star of node 5, we obtain the graph of Fig. 1(b) and a facet-defining inequality; the coefficients different from 1 appear in the figure. The right-hand side is 7. Again, if we apply Theorem 2.3 to the stars of nodes 1, 2, 3, 4, and 5 in Fig. 1(a), we obtain a facet-defining inequality whose right-hand side is 12 and whose coefficients different from 1 appear in Fig. 1(c). Finally, if we apply Theorem 2.3 to the star of 6 in Fig. 1(a) and then Theorem 2.4 to subdivide some edges, we also obtain the graph in Fig. 1(c) but a different inequality whose right-hand side is 10 and whose coefficient different from 1 appears in Fig. 1(d).

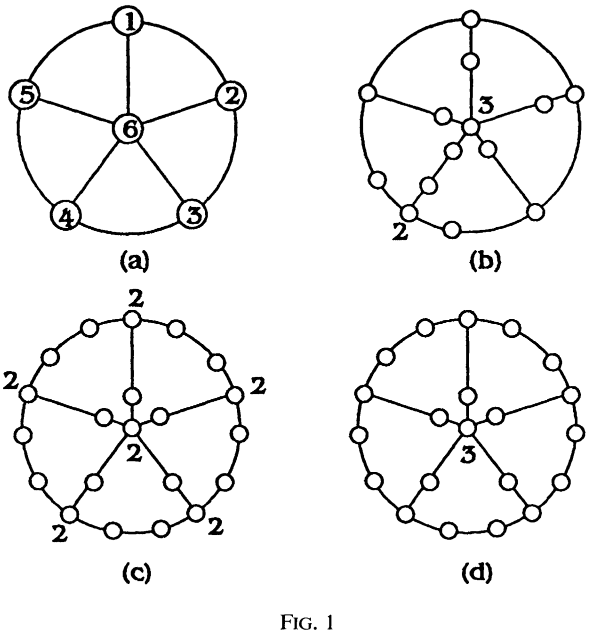


FIG. 1

3. Compositions of graphs. Let $G = (V, E)$ be a graph such that $V = V_1 \cup V_2$, $W = V_1 \cap V_2 \neq \emptyset$ and $(W, E(W))$ is a clique and $(V \backslash W, E(V \backslash W))$ is disconnected. Chvátal [5] proved the following.

THEOREM 3.1. *If $G_1 = (V_1, E(V_1))$, $G_2 = (V_2, E(V_2))$, then a system that defines $P(G)$ is obtained by taking the union of the systems that define $P(G_1)$ and $P(G_2)$.*

This theorem applies to the case where G has a one-node cutset or a two-node cutset $\{u, v\}$ with $uv \in E$; we refer to this as Case 1.

In the remainder of this section, we assume that

(i) $V = V_1 \cup V_2$,
(ii) $V_1 \cap V_2 = \{u, v\}$,
(iii) $G \backslash \{u, v\}$ is disconnected,
(iv) The nodes u and v are not adjacent.

This will be called Case 2. We add a five-cycle to each piece. We shall see that we can easily derive a description of the polytope for the original graph from the polytopes of the modified pieces. Let $\bar{G}_k = (\bar{V}_k, \bar{E}_k)$ be defined as follows:

(i) $\bar{V}_k = V_k \cup \{w_1, w_2, w_3\}$,
(ii) $\bar{E}_k = E(V_k) \cup \{uw_1, vw_1, uw_2, w_2w_3, w_3v\}$ for $k = 1, 2$. Let $\bar{G} = (\bar{V}, \bar{E})$ be the union of $\bar{G}_1$ and $\bar{G}_2$, i.e., $\bar{V} = \bar{V}_1 \cup \bar{V}_2$, $\bar{E} = \bar{E}_1 \cup \bar{E}_2$.

The inequality

$$\sum_{i=1}^{3} x(w_i) + x(u) + x(v) \leq 2 \tag{3.1}$$

defines a facet $F(\bar{G}_k)$ of $P(\bar{G}_k)$, $k = 1, 2$ and a facet $F(\bar{G})$ of $P(\bar{G})$. Furthermore, the polytope $P(G)$ is the projection of $F(\bar{G})$ along the variables $\{x(w_i)\}$, i.e.,

$$P(G) = \{y \mid (y, x(w_1), x(w_2), x(w_3))^t \in F(\bar{G})\}.$$

The next lemma gives a system that defines $F(\bar{G})$.

LEMMA 3.2. *Given two systems of inequalities defining $F(\bar{G}_1)$ and $F(\bar{G}_2)$, the union of these two systems defines $F(\bar{G})$.*

Proof. See Theorem 2.4 in [2]. □

Lemmas 2.1 and 2.2 show that the facet-defining inequalities of $P(\bar{G}_k)$ can be classified as follows, for $k = 1, 2$:

$$\sum_{j \in V_k} a_{ij}^k x(j) \leq \alpha_i^k, \qquad i \in I_1^k, \tag{3.2a}$$

$$\sum_{j \in V_k} a_{ij}^k x(j) + x(w_1) \leq \alpha_i^k, \qquad i \in I_2^k, \tag{3.2b}$$

$$\sum_{j \in V_k} a_{ij}^k x(j) + x(w_2) + x(w_3) \leq \alpha_i^k, \qquad i \in I_3^k, \tag{3.2c}$$

$$x(u) + x(w_1) \leq 1, \tag{3.2d}$$

$$x(u) + x(w_2) \leq 1, \tag{3.2e}$$

$$x(v) + x(w_1) \leq 1, \tag{3.2f}$$

$$x(v) + x(w_3) \leq 1, \tag{3.2g}$$

$$x(w_2) + x(w_3) \leq 1, \tag{3.2h}$$

$$x(u) + x(v) + x(w_1) + x(w_2) + x(w_3) \leq 2, \tag{3.2i}$$

$$x(j) \geq 0, \qquad j \in \bar{V}_k. \tag{3.2j}$$

The set I_1^k consists of the inequalities whose support does not intersect $\{w_1, w_2, w_3\}$. The set I_2^k contains the inequalities whose support includes $\{u, v, w_1\}$ and has empty intersection with $\{w_2, w_3\}$. The inequalities in I_3^k have a support that contains $\{u, v, w_2, w_3\}$ and does not include w_1.

Then $F(\bar{G})$ is defined by both systems together with the inequality

$$-x(u) - x(v) - x(w_1) - x(w_2) - x(w_3) \leq -2. \tag{3.3}$$

Now we project the variables $\{x(w_i)\}$ using the following result of Balas and Pulleyblank [1].

THEOREM 3.3. *Let* $Z = \{(w, x) | Aw + Bx \leq b, w \geq 0, x \geq 0\}$ *the projection of* Z *along the subspace of the* w *variables is*

$$X = \{x | (vB)x \leq vb, \forall v \in \text{extr } \Psi, x \geq 0\},$$

where extr Ψ *denotes the set of extreme rays of*

$$\Psi = \{y | yA \geq 0, y \geq 0\}.$$

In our case, the rows of A are of the following types:

$$\begin{array}{ccc} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \end{array}.$$

The first nine rows correspond to inequalities (3.1a)–(3.1i), and the last row corresponds to (3.3). The extreme rays of Ψ correspond to the extreme points of

$$\{y | yA \geq 0, \sum y_i = 1, y \geq 0\},$$

so we enumerate the extreme points of

$$\{z | Bz \geq 0, \sum z_i = 1, z \geq 0\},$$

where B is the matrix

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 1 & 1 & -1 \end{bmatrix}.$$

The extreme points are the columns of the matrix below:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & & \vdots & \frac{1}{3} & \frac{1}{4} & 0 \\ & 0 & \ddots & & \frac{1}{3} & 0 & 0 \\ & & & & 0 & \frac{1}{4} & 0 \\ \vdots & \vdots & & & 0 & \frac{1}{4} & 0 \\ & & & 1 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \cdots & 0 & \frac{1}{3} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Therefore the inequalities given by Theorem 3.3 are obtained by performing the following steps:

(i) Keep inequalities (3.2a)–(3.2i) but delete the variables $\{x(w_i)\}$,

(ii) Add three inequalities, one of type (3.2b), (3.2d), or (3.2f), one of type (3.2c) or (3.2h), and (3.3),

(iii) Add four inequalities, one of type (3.2b), (3.2d), or (3.2f), one of type (3.2e), one of type (3.2g), and (3.3),

(iv) Add (3.2i) and (3.3) (this gives the redundant inequality $0 \leq 0$).

The next lemma shows that some of those inequalities are redundant.

LEMMA 3.4. *Let $ax \leq \alpha$ be an inequality that defines a nontrivial facet of $P(G)$. If $V_a \subseteq V_k$, then this inequality also defines a facet of $P(\bar{G}_k)$, $k = 1, 2$.*

Proof. Let H be the graph obtained by replacing the edge uw_1 in $\bar{G}_k$ by the path us, st, tw_1, where s and t are new nodes. First, we prove that $ax \leq \alpha$ defines a facet of $P(H)$. It is clear that this inequality is valid for $P(H)$.

By hypothesis, there are $|V_k|$ linearly independent incidence vectors of stable sets of G_k that satisfy $ax = \alpha$. Let us form a matrix M with them. Among these vectors, there is one vector $\bar{x}$ such that $\bar{x}(u) = 0$ and one vector $\tilde{x}$ such that $\tilde{x}(v) = 0$. Consider the matrix

$$M' = \begin{pmatrix} M & \bar{x} & \bar{x} & \tilde{x} & \bar{x} & \tilde{x} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The last five rows correspond to s, t, w_1, w_2, and w_3, respectively.

The columns of M' are linearly independent incidence vectors of stable sets of H and satisfy $ax = \alpha$. Now, by Theorem 2.5, we can replace the path us, st, tw_1 by the edge uw_1 and we have that $ax \leq \alpha$ defines a facet of the stable set polytope of $\bar{G}_k$. □

This lemma states that, if an inequality is essential in the definition of $P(G)$ and its support is included in V_k, then a multiple of this inequality already appears in (3.2a). So when we apply (i) we should keep only the inequality (3.2a); in (ii) we should keep only those constraints whose support intersects both $V_1 \backslash \{u, v\}$ and $V_2 \backslash \{u, v\}$; for the same reasons the inequalities produced in (iii) are redundant. We can state the main result of this paper.

THEOREM 3.5. *The polytope $P(G)$ is defined by* (3.2a), *together with $x(j) \geq 0$, for $j \in V$, and the mixed inequalities*

$$\sum_{j \in V_k} a_{ij}^k x(j) + \sum_{j \in V_l} a_{sj}^l x(j) - x(u) - x(v) \leq \alpha_i^k + \alpha_s^l - 2 \tag{3.4}$$

$$\textit{for } k = 1, 2;\ l = 1, 2;\ k \neq l;\ i \in I_2^k,\ s \in I_3^l.$$

To prove that this system is minimal, we introduce the following lemma.

LEMMA 3.6. *Inequalities* (3.2b) *and* (3.2c) *define facets of $F(\bar{G}_k)$.*

Proof. First, we study inequality (3.2b). Let $ax \leq \alpha$ be one of them and let $S = \{x_1, \ldots, x_p\}$ be the set of extreme points of $P(\bar{G}_k)$ that satisfy $ax = \alpha$.

Let x_j be a vector in S; if $x_j \in F(\bar{G}_k)$, we define $x_j' = x_j$; otherwise, we define $x_j'(i) = x_j(i)$ if $i \neq w_2, w_3$. We set to 1 $x_j'(w_2)$ or $x_j'(w_3)$ to obtain an extreme point $x_j' \in F(\bar{G}_k)$. Now let us assume that $ax \leq \alpha$ is inessential in the definition of $F(\bar{G}_k)$. This implies that $a = \sum \lambda_i b_i + \gamma d$, $\alpha \geq \sum \lambda_i \beta_i + \gamma \varepsilon$, with $\lambda_i \geq 0$, where $b_i x \leq \beta_i$ denotes an inequality of (3.2) different from $ax \leq \alpha$ and from (3.2i); $dx = \varepsilon$ denotes the equation obtained from (3.2i).

For each inequality $b_i x \leq \beta_i$, there is a vector $x_j \in S$ such that $b_i x_j < \beta_i$ and $b_i x'_j = \beta_i$. Hence $b_{iw_j} > 0$ for $j = 2$ or 3; then $b_{iw_1} = 0$. This implies $\gamma = 1$ and then some $\lambda_i < 0$, a contradiction.

Now we consider an inequality $ax \leq \alpha$ of the type (3.2c).

Since dim $(F(\bar{G}_k)) = |\bar{V}_k| - 1$, we must show $|\bar{V}_k| - 1$ linearly independent vectors in $F(\bar{G}_k)$ that satisfy $ax = \alpha$. Let $S = \{x_1, \ldots, x_p\}$ be the set of extreme points of $P(\bar{G}_k)$ that satisfy $ax = \alpha$. Let x_j be a vector in S; if $x_j \in F(\bar{G}_k)$, we set $x'_j = x_j$; otherwise, we set $x'_j(i) = x_j(i)$, if $i \neq w_1$, $x'_j(w_1) = 1$. Since S contains $|\bar{V}_k|$ linearly independent vectors and we modified only one component of the vectors in S, the set $S' = \{x'_1, \ldots, x'_p\}$ contains $|\bar{V}_k| - 1$ linearly independent vectors in $F(\bar{G}_k)$ that satisfy $ax = \alpha$. □

So if $ax \leq \alpha$ is of type (3.2b) or (3.2c) and another inequality $bx \leq \beta$ defines the same face of $F(\bar{G}_k)$, then $b = \lambda a + \gamma d$, where $dx = \varepsilon$ denotes (3.2i) and $\lambda \geq 0$. The constraint $bx \leq \beta$ cannot be in (3.2) because of the structure of the inequalities in this system.

COROLLARY 3.7. *Inequalities* (3.4) *define facets of* $P(G)$.

Proof. Assume that $k = 1$, $l = 2$. We show that there exists a vector $\bar{x} \in P(G)$ that satisfies this inequality and all others as strict inequalities.

For the inequality

$$\sum_{j \in V_1} a^1_{ij} x(j) + x(w_1) \leq \alpha^1_i, \qquad i \in I^1_2,$$

let $S = \{x_1, \ldots, x_n\}$ be the set of extreme points of $F(\bar{G}_1)$ that satisfy it. For

$$\sum_{j \in V_2} a^2_{ij} x(j) + x(w_2) + x(w_3) \leq \alpha^2_i, \qquad i \in I^2_3,$$

let $T = \{y_1, \ldots, y_m\}$ be the set of extreme points of $F(\bar{G}_2)$ that satisfy it. First, for each vector x_i, we find a vector y_{j_i} such that together they give a vector $z_i \in F(\bar{G})$, $i = 1, \ldots, n$. Similarly, for each vector y_i, we find a vector x_{j_i} that gives a vector $z_{n+i} \in F(\bar{G})$, $i = 1, \ldots, m$. Let $\{z_1, \ldots, z_r\}$ be the set of vectors thus obtained, $(r = n + m)$. Let z'_k be the vector obtained by dropping the components $z_k(w_j)$, $1 \leq j \leq 3$ from z_k; then

$$\bar{x} = \frac{1}{r}(z'_1 + \cdots + z'_r)$$

is the required vector. □

4. Algorithmic aspects. The optimization problem can be also decomposed. The following algorithm appeared in Boulala and Uhry [4] and Sbihi and Uhry [12].

Let $G = (V, E)$ be a graph and $c : V \to \Re_+$ a weight function. Let us assume that G is the graph of Theorem 3.1, let $W = \{w_1, \ldots, w_l\}$, and let β_i be the maximum weight of a stable set of G_2 that contains w_i for $1 \leq i \leq l$. Let β_0 be the maximum weight of a stable set of G_2 that does not contain any node of W. Let us redefine the weights in G_1 as follows:

$$c'(u) = c(u) \qquad \text{if } u \notin W,$$

$$c'(w_i) = \max\{0, \beta_i - \beta_0\} \quad \text{for } 1 \leq i \leq l.$$

Let α be the maximum weight of a stable set of G_1; then the maximum weight of a stable set of G is $\alpha + \beta_0$.

Now let us study Case 2 of §3. Let $y_0 = u$, $y_1 = w_2$, $y_2 = w_3$, $y_3 = v$, $y_4 = w_1$. For $0 \le i \le 4$, let β_i be the maximum weight of a stable set of $\bar{G}_2$ whose node set contains y_i and y_{i+2} (indices taken mod 5); the weights of the nodes $\{w_i\}$ are zero.

Let $[\gamma_0, \dots, \gamma_4]$ be the solution of the system

$$(\gamma_0, \dots, \gamma_4)\begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4).$$

We have that

$$\gamma_0 = \frac{\beta_0 - \beta_1 - \beta_2 + \beta_3 + \beta_4}{2},$$

$$\gamma_1 = \frac{\beta_0 + \beta_1 - \beta_2 - \beta_3 + \beta_4}{2},$$

$$\gamma_2 = \frac{\beta_0 + \beta_1 + \beta_2 - \beta_3 - \beta_4}{2},$$

$$\gamma_3 = \frac{-\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4}{2},$$

$$\gamma_4 = \frac{-\beta_0 - \beta_1 + \beta_2 + \beta_3 + \beta_4}{2}.$$

Let $M = \min\{\gamma_i\}$; if the numbers $\{\beta_i\}$ are integers, then $\gamma_i - M$ is a nonnegative integer for $0 \le i \le 4$. Now we define

$$c'(j) = c(j) \quad \text{if } j \in V_1 \backslash \{u, v\},$$

$$c'(y_i) = \gamma_i - M \quad \text{for } 0 \le i \le 4.$$

Let α be the maximum weight of a stable set of $\bar{G}_1$; then the maximum weight of a stable set of G is $\alpha + 2M$. Let us remark that one of the weights $c'(y_k)$ is zero; we can then delete that node from $\bar{G}_1$.

5. Application to series-parallel graphs.

Boulala and Uhry [4] proved the following.

THEOREM 5.1. *If $G = (V, E)$ is a series-parallel graph, then $P(G)$ is defined by*

$$(5.1)\qquad \begin{aligned} &0 \le x(u) \quad \textit{for all } u \in V, \\ &x(u) + x(v) \le 1 \quad \textit{for all } uv \in E, \\ &\sum_{u \in C} x(u) \le \frac{|C| - 1}{2} \quad \textit{for all odd holes } C. \end{aligned}$$

They also gave a linear time algorithm to find a maximum weighted stable set in a series-parallel graph. A short proof of Theorem 5.1 appears in [9].

A connected series-parallel graph can be decomposed into paths and triangles using one-node and two-node cutsets. If G consists of a cycle and a path joining two nodes of the cycle, then it is easily seen that $P(G)$ is defined by (5.2). Thus Theorem 5.1 can be derived from Theorems 3.1 and 3.5.

Since a series-parallel graph may have exponentially many odd cycles, the stable set polytope may have exponentially many facets.

Hassin and Tamir [7] proved that, if $G = (V, E)$ is a series-parallel graph, then it contains a two-node cutset such that, when the decomposition of §3 is carried out, $|V_i| \leq 2/3|V| + 2$ for $i = 1, 2$. Hence we can recursively decompose a series-parallel graph until each piece has at most fifteen nodes. By Lemma 3.2, we can describe a polytope Q such that $P(G)$ is a projection of Q. If G has n nodes, then the number of inequalities and the number of variables in the system that defines Q is $O(n)$. Such a system is compact.

6. Graphs with no $K_5\backslash e$ minor. A graph G is said to contain a graph H as a minor if a graph isomorphic to H can be obtained from G by repeated deletion and contraction of edges of G. Let us denote by $\mathscr{C}$ the class of all connected graphs that do not contain $K_5\backslash e$ as a minor; this is the graph K_5 minus one edge. Wagner [16] gave a characterization of the graphs in $\mathscr{C}$.

If G_1 and G_2 are node-disjoint graphs with at least two nodes, v_1 a node of G_1 and v_2 a node of G_2 then the 1-sum of G_1 and G_2 (with respect to v_1 and v_2) is obtained by identifying the nodes v_1 and v_2. If e_1 is an edge of G_1 and e_2 is an edge of G_2, then the 2-sum of G_1 and G_2 (with respect to e_1 and e_2) is obtained by identifying e_1 and e_2 (and, of course, the endnodes of e_1 and e_2).

Wagner proved that each maximal graph G in $\mathscr{C}$ (i.e., by adding a further edge to G, the new graph will contain $K_5\backslash e$ as a minor) can be obtained by starting with the graphs of Fig. 2 and taking repeated 1-sums or 2-sums. Equivalently, if we have a maximal graph $G \in \mathscr{C}$, we can decompose it into the graphs of Fig. 2. If the graph $G \in \mathscr{C}$ is not maximal, then we may use also spanning subgraphs of the graphs in Fig. 2. In this case, the nodes of the two-node cutset may be nonadjacent in G.

Let n be the number of nodes of $G \in \mathscr{C}$; a two-vertex cutset can be found in $O(n)$ time. The pieces that we obtain are the graphs of Fig. 2, where some edges are replaced by a five-cycle. The max stable set problem in a graph of this type obtained from a wheel can be solved in linear time, so, using the procedure of §4, we can find a maximum weighted stable set in G in $O(n^2)$ time.

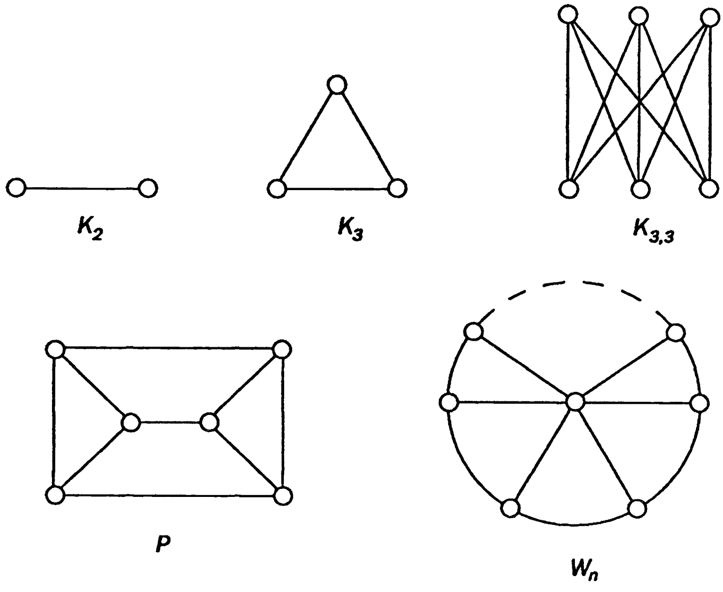
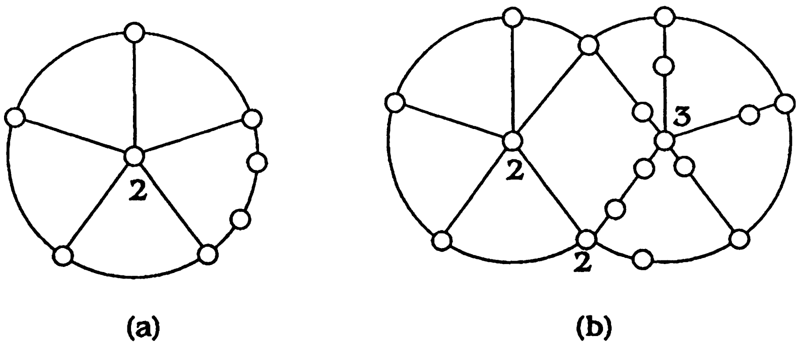


FIG. 2



(a) (b)

FIG. 3

The development of a polynomial algorithm for combinatorial optimization problems has often been closely related to the characterization of a system of linear inequalities that defines the corresponding polytope. This is the case for the stable set polytope of series-parallel graphs [4]. The missing piece here is a characterization of the polytope for subdivisions of wheels; in §2 we gave procedures to produce some of these inequalities. In [3] we used a tour de force to characterize the polytope for subdivisions of K_4. In what follows, we present some examples of facet-defining inequalities of $P(G)$ for $G \in \mathscr{C}$.

The graph of Fig. 3(a) has been obtained by subdividing one edge of the graph in Fig. 1(a). Theorem 2.4 gives us a facet-defining inequality whose right-hand side is 3 and whose coefficient different from 1 is shown in the figure. We can compose this graph with the graph of Fig. 1(b) to obtain the graph in Fig. 3(b). By Corollary 3.7, we can see that there is facet-defining inequality whose right-hand side is 8 and whose coefficients different from 1 appear in the figure. Given any positive integer p, we can construct a graph $G \in \mathscr{C}$ and an inequality that defines a facet of $P(G)$ with coefficients $1, 2, \dots, p$. For this, it is enough to compose subdivisions of wheels of different sizes.

7. Composition of *h*-perfect graphs. A graph G is said to be h-perfect if $P(G)$ is defined by the constraints corresponding to cliques, odd holes, and the nonnegativity constraints.

Let G be the graph of Theorem 3.5; we can derive h-perfectness of G as follows:

(a) If $\bar{G}_1$ and $\bar{G}_2$ are h-perfect then G is also h-perfect;

(b) If $\bar{G}_2\backslash\{w_1\}$ is h-perfect, $\bar{G}_1$ is h-perfect and the set of inequalities (3.2c) for $P(\bar{G}_1)$ is empty then G is h-perfect;

(c) If $\bar{G}_2\backslash\{w_2, w_3\}$ is h-perfect, $\bar{G}_1$ is h-perfect and the set of inequalities (3.2b) for $P(\bar{G}_1)$ is empty then G is h-perfect.

Sbihi and Uhry [12] studied graphs G, which are the union of a bipartite graph G_1 and graph G_2 having exactly two common nodes u and v and no edge in common. They proved that the graph G is h-perfect if the graph obtained from G by replacing G_1 by an $u - v$ chain is h-perfect. They also proved that the graph obtained by substituting bipartite graphs for edges of a series parallel graph is h-perfect. Their results follow from remarks (a)–(c).

Gerards [6] studied graphs with not odd K_4. Those graphs can be decomposed by two-node cutsets as shown by Lovász et al. [8]. Gerards used compositions similar to those of Boulala and Uhry [4] and Sbihi and Uhry [12].

Acknowledgment. We thank the referee for useful suggestions regarding the presentation of this work.

REFERENCES

[1] E. BALAS AND W. R. PULLEYBLANK, *The perfectly matchable subgraph polytope of a bipartite graph*, Networks, 13 (1983), pp. 495–516.

[2] F. BARAHONA AND A. R. MAHJOUB, *Compositions of graphs and polyhedra* I: *Balanced induced subgraphs and acyclic subgraphs*, SIAM J. Discrete Math., 7 (1994), pp. 344–358, this issue.

[3] ———, *Compositions of Graphs and Polyhedra* III: *Graphs with no W_4 minor*, SIAM J. Discrete Math., 7 (1994), pp. 372–389, this issue.

[4] M. BOULALA AND J. P. UHRY, *Polytope des indepéndants d'un graphe série-parallèle*, Discrete Math., 27 (1979), pp. 225–243.

[5] V. CHVÁTAL, *On certain polytopes associated with graphs*, J. Combin. Theory Ser. B, 18 (1975), pp. 138–154.

[6] A. M. H. GERARDS, *An extension of König's Theorem to graphs with no odd K_4*, Tilburg University, 1987.

[7] R. HASSIN AND A. TAMIR, *Efficient algorithms for optimization and selection on series-parallel graphs*, SIAM J. Discrete Algebraic Meth., 7 (1986), pp. 379–389.

[8] L. LOVÁSZ, A. SCHRIJVER, P. D. SEYMOUR, AND K. TRUEMPER, unpublished paper, 1984.

[9] A. R. MAHJOUB, *On the stable set polytope of a series-parallel graph*, Math. Programming, 40 (1988), pp. 53–57.

[10] G. L. NEMHAUSER AND L. E. TROTTER, *Properties of vertex packing and independence system polyhedra*, Math. Programming, 8 (1975), pp. 232–248.

[11] M. W. PADBERG, *On the facial structure of set packing polyhedra*, Math. Programming, 5 (1973), pp. 199–215.

[12] N. SBIHI AND J. P. UHRY, *A class of h-perfect graphs*, Discrete Math., 51 (1984), pp. 191–205.

[13] L. E. TROTTER, *A class of facet producing graphs for vertex packing polyhedra*, Discrete Math., 12 (1975), pp. 373–388.

[14] L. E. TROTTER AND R. GILES, *On stable set polyhedra for $K_{1,3}$-free graphs*, J. Combin. Theory Ser. B, 31 (1981), pp. 313–326.

[15] L. WOLSEY, *Further facet generating procedures for vertex packing polytopes*, Math. Programming, 11 (1979), pp. 158–163.

[16] K. WAGNER, *Bemerkungen zu Hadwigers Vermatung*, Math. Ann., 141 (1960), pp. 433–451.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 372–389, August 1994

COMPOSITIONS OF GRAPHS AND POLYHEDRA III: GRAPHS WITH NO W_4 MINOR*

FRANCISCO BARAHONA[†] AND ALI RIDHA MAHJOUB[‡]

Abstract. The authors characterize the stable set polytope for graphs that do not have a 4-wheel as a minor. The authors prove that the nontrivial facets are either "edge" inequalities or can be obtained by composing "odd cycles" and "subdivisions of K_4." By adding some extra variables, it is shown that the stable set problem for these graphs can be formulated as a linear program of polynomial size.

Key words. polyhedral combinatorics, composition of polyhedra, stable set polytope, compact systems

AMS subject classifications. 05C85, 90C27

1. Introduction.

Given a graph $G = (V, E)$, a set $S \subseteq V$ is called a *stable set* if no two nodes in S are adjacent. Given a stable set S, the incidence vector of S, $x^S \in \Re^V$ is defined by

$$x^S(u) = \begin{cases} 1 & \text{if } u \in S, \\ 0 & \text{if } u \notin S. \end{cases}$$

The *stable set polytope* of G, denoted by $P(G)$, is the convex hull of incidence vectors of stable sets of G. The maximum stable set problem is NP-hard, so it seems difficult to find a complete characterization of $P(G)$ for general graphs. To our knowledge, the only classes of graphs, besides perfect graphs, for which this polytope has been characterized are line graphs [4], series parallel graphs [2], [13], almost bipartite graphs [5], and graphs with no odd K_4 [7]. The class studied by Gerards and Schrijver [7] contains the classes studied by Boulala, Fonlupt, and Uhry. In this case, the only nontrivial facets correspond to edges and odd holes. All the linear systems mentioned above consist of inequalities with 0-1 coefficients.

In this paper, we characterize the stable set polytope for graphs that do not have a 4-wheel as a minor. The inequalities are more difficult to describe than in the preceding cases, and they may have arbitrarily large coefficients.

Graphs in this class can be decomposed by two-vertex cuts [10]. We use this property to prove that the nontrivial facets are either edges or can be obtained by composing odd cycles and subdivisions of K_4. A list of the facets for subdivisions of K_4 is also given. We also show that, by adding some extra variables, the stable set problem in graphs with no 4-wheel minor can be formulated as a linear program of polynomial size. A polynomial combinatorial algorithm for the stable set problem in this class can be easily derived [1].

If G has a one-node or a two-node cutset, then G decomposes into G_1 and G_2. In a companion paper [1], we gave a technique to characterize $P(G)$ starting from systems related to G_1 and G_2. In this paper, we apply that technique.

* Received by the editors June 4, 1990; accepted for publication (in revised form) May 20, 1993. Part of this work has been done at the Institute for Operations Research of the University of Bonn and supported by the joint project "Combinatorial Optimization" of the Natural Sciences and Engineering Research Council of Canada and the German Research Association (Deutsche Forschungsgemeinschaft, SFB 303).

[†] IBM T. J. Watson Research Center, Yorktown Heights, New York 10598.

[‡] Department of Statistics, King Saud University, Riyadh, Saudi Arabia. Present address, Laboratoire d'Informatique de Brest (LIBr), Université de Bretagne Occidentale, 6 Avenue Le Gorgeu, 29287 Brest Cedex, France.

A connected graph G is said to have the graph H as a *minor* if H can be obtained from G by deleting some edges and by a sequence of elementary contractions in which a pair of adjacent vertices is identified and all other adjacencies between vertices are preserved (multiple edges arising from the identification being replaced by single edges).

We denote a 4-wheel by W_4; see Fig. 1.1.

This paper is organized as follows. In §2 we summarize our composition techniques. In §3 we describe the decomposition of graphs with no W_4 minor. In §4 we state our main result. Sections 5 and 6 are devoted to the study of the subdivisions of K_4.

We conclude this introduction with a few definitions.

The polytope $P(G)$ is full-dimensional. This implies that (up to multiplication by a positive constant) there is a unique nonredundant inequality system $Ax \leq b$ such that $P(G) = \{x : Ax \leq b\}$. These inequalities define the *facets* of $P(G)$. In many cases, we say that the inequality $ax \leq \alpha$ *is a facet* instead of saying that it defines a facet. If the inequality has at least two nonzero coefficients, we say that it is a *nontrivial* facet.

If $ax \leq \alpha$ defines a facet of $P(G)$, we denote by V_a the set

$$V_a = \{v : a_v > 0\}.$$

The subgraph induced by V_a is denoted by G_a, and it is called the *support* of the facet.

We denote by uv the edge whose endnodes are u and v. If $U \subseteq V$, then $E(U)$ denotes the set of edges with both endnodes in U, and $(U, E(U))$ is the *subgraph induced* by U. An odd cycle with no chord is called an *odd hole*. A maximal complete graph is called a *clique*.

If K is a clique, then the inequality $\sum_{u \in K} x(u) \leq 1$ defines a facet of $P(G)$ [14]. This is called a *clique inequality*. If the clique is an edge, it is called an *edge inequality*.

If H is an odd hole, then the inequality

$$\sum_{u \in H} x(u) \leq \frac{|H| - 1}{2}$$

is valid for $P(G)$; this is called an *odd hole inequality*. Under some conditions, these inequalities define facets of $P(G)$ [14].

The trivial facets of $P(G)$ are $x(v) \geq 0$ for $v \in V$.

A graph G is called *t-perfect* if the only nontrivial facets of $P(G)$ are the odd hole and the edge inequalities. Chvátal [3] introduced this class of graphs and conjectured that series-parallel graphs are t-perfect.

A graph is called *series-parallel* if it does not contain K_4 as a minor. Boulala and Uhry [2] proved that series-parallel graphs are t-perfect; i.e., they characterized $P(G)$ for graphs that do not have a 3-wheel as a minor. A short proof of this appears in [13].

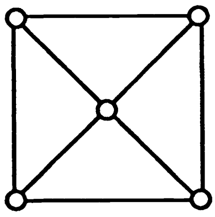


FIG. 1.1

2. Compositions of polyhedra. This section is devoted to survey the composition/decomposition techniques that we need.

Let $G = (V, E)$ be a graph such that $V = V_1 \cup V_2$, $W = V_1 \cap V_2 \neq \emptyset$, $(W, E(W))$ is a clique, and $(V \backslash W, E(V \backslash W))$ is disconnected. Chvátal [3] proved the following result.

THEOREM 2.1. *If $G_1 = (V_1, E(V_1))$ and $G_2 = (V_2, E(V_2))$, then a system that defines $P(G)$ is obtained by taking the union of the systems that define $P(G_1)$ and $P(G_2)$ and identifying the variables associated with the nodes in W.*

This theorem applies to the case where G has a one-node or a two-node cutset $\{u, v\}$ with $uv \in E$. This is called Case 1.

Now we must treat Case 2, i.e., when G has a two-node cutset $\{u, v\}$ and $uv \notin E$.

In the remainder of this section, we assume that

(i) $V = V_1 \cup V_2$,

(ii) $V_1 \cap V_2 = \{u, v\}$,

(iii) $uv \notin E$, and

(iv) $G \backslash \{u, v\}$ is disconnected.

We decompose into two pieces and add a 5-cycle to both of them; see Fig. 2.1. For $k = 1, 2$, we define $\bar{G}_k = (\bar{V}_k, \bar{E}_k)$ as

(i) $\bar{V}_k = V_k \cup \{w_1, w_2, w_3\}$,

(ii) $\bar{E}_k = E(V_k) \cup \{uw_1, vw_1, uw_2, w_2w_3, w_3v\}$.

To study the facets of $P(\bar{G}_k)$, we present two lemmas. Their proofs appear in [13].

LEMMA 2.2. *Let $ax \leq \alpha$ be a facet of $P(G)$. If G_a has a path with vertices p, u, v, q, where u and v have degree 2 in G_a, then $a_u = a_v$.*

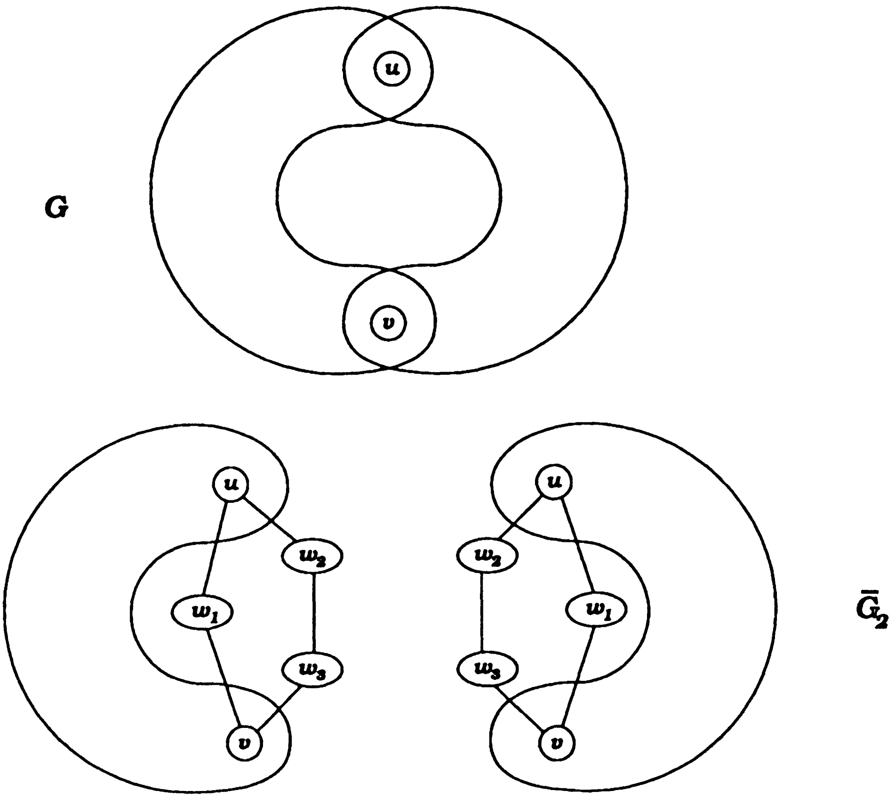


FIG. 2.1

LEMMA 2.3. *Let $ax \leq \alpha$ be a facet of $P(G)$. If G_a is different from an odd hole, then it does not contain between two given nodes p and q two paths such that each node of them, different from p, q, has degree 2 in G_a.*

Lemmas 2.2 and 2.3 imply that the facets of $P(\bar{G}_k)$ for $k = 1, 2$ can be classified in the following ten types:

(a) $\sum_{j\in V_k} a^k_{ij}x(j) \leq \alpha^k_i$, $i \in I^k_1$,
(b) $\sum_{j\in V_k} a^k_{ij}x(j) + x(w_1) \leq \alpha^k_i$, $i \in I^k_2$,
(c) $\sum_{j\in V_k} a^k_{ij}x(j) + x(w_2) + x(w_3) \leq \alpha^k_i$, $i \in I^k_3$,
(d) $x(u) + x(w_1) \leq 1$,
(e) $x(u) + x(w_2) \leq 1$,
(f) $x(v) + x(w_1) \leq 1$,
(g) $x(v) + x(w_3) \leq 1$,
(h) $x(w_2) + x(w_3) \leq 1$,
(i) $x(u) + x(v) + x(w_1) + x(w_2) + x(w_3) \leq 2$,
(j) $x(j) \geq 0, j \in \bar{V}_k$,

where I^k_1 is the set of inequalities whose support has empty intersection with $\{w_1, w_2, w_3\}$, I^k_2 is the set of inequalities whose support contains w_1 and has empty intersection with $\{w_2, w_3\}$, and I^k_3 is the set of inequalities whose support contains $\{w_2, w_3\}$ and not w_1.

Now we can present the necessary polyhedral composition theorems.

Let $\bar{G} = (\bar{V}, \bar{E})$ be the union of $\bar{G}_1$ and $\bar{G}_2$, i.e.,

$$\bar{V} = \bar{V}_1 \cup \bar{V}_2, \qquad \bar{E} = \bar{E}_1 \cup \bar{E}_2.$$

The equation

$$x(u) + x(v) + x(w_1) + x(w_2) + x(w_3) = 2 \tag{2.1}$$

defines a facet $F(\bar{G})$ of $P(\bar{G})$; it also defines a facet $F(\bar{G}_k)$ of $P(\bar{G}_k)$ for $k = 1, 2$. The polytope $P(G)$ is a projection of $F(\bar{G})$ along the variables $\{x(w_i)\}$.

Now we state two theorems that appear in [1].

THEOREM 2.4. *The facet $F(\bar{G})$ is defined by the union of the systems that define $F(\bar{G}_1)$ and $F(\bar{G}_2)$.*

THEOREM 2.5. *The polytope $P(G)$ is defined by* (a), *together with $x(j) \geq 0$ and the mixed inequalities*

$$\sum_{j\in V_k} a^k_{ij}x(j) + \sum_{j\in V_l} a^l_{sj}x(j) - x(u) - x(v) \leq \alpha^k_i + \alpha^l_s - 2 \tag{2.2}$$

$$\text{for } k = 1, 2;\ l = 1, 2;\ k \neq l;\ i \in I^k_2,\ s \in I^l_3.$$

Moreover, all these inequalities define facets of $P(G)$.

3. Graphs with no W_4 minor.

Graphs with no W_4 minor can be easily decomposed [10]. Gan and Johnson [6] used this property to study the Chinese postman problem in these graphs. More precisely, if G has no W_4 as a minor and has at least five nodes, then G has a one-node or a two-node cutset where one of the pieces is a path or the *Wheatstone bridge*; see Fig. 3.1.

Now it is clear how to apply the decomposition techniques of §2. If the cutset is $\{u, v\}$ and $uv \in E$, then we just separate the two pieces. If $uv \notin E$, then we separate the two pieces and add a 5-cycle to both of them. In what follows, we formalize this procedure. Let us denote by n the number of nodes of G; we prove that the total number of nodes after decomposing is $O(n)$.

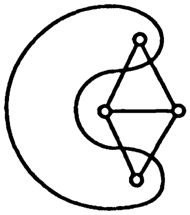


FIG. 3.1

We recursively apply the procedure below.

(a) If G has at most four nodes, stop.

(b) If G has a one-node cutset, we decompose it into the two blocks.

(c) Suppose that G has a two-node cutset $\{u, v\}$, where the second block is a path with two edges or the Wheatstone bridge.

(i) If $uv \notin E$, we decompose G into the two blocks, we add the edge uv to both blocks, and we label these two new edges as "artificial." This corresponds to Case 2 of §2. Artificial edges represent the 5-cycles that are added to both pieces.

(ii) If $uv \in E$ (where uv is not artificial), we decompose into the two blocks. This corresponds to Case 1 of §2.

(iii) If $uv \in E$ and uv is artificial, we decompose into the two blocks, we leave the artificial edge uv only in the first block (if there are parallel artificial edges between u and v, we leave them all in the first block), and we add a new artificial edge uv to each block. This corresponds to Case 2 of §2.

Note that a two vertex cutset could be used several times in this decomposition and that that would create parallel artificial edges. The number of nodes of the larger block decreases each time we decompose, so the number of artificial edges is bounded by $2n$. The resulting pieces are single edges, sets of parallel edges, triangles, or copies of K_4. Therefore, after applying this procedure, the total number of edges is $O(n)$. These pieces may have parallel artificial edges.

Figure 3.2 shows an example of this decomposition. Dashed lines represent artificial edges. The set $\{u, v\}$ has been used twice in the decomposition.

Now we must treat the blocks that have parallel artificial edges. Given a block with more than two nodes and parallel artificial edges between u and v, we decompose into two blocks. One of them consists of all those parallel edges. We add a new artificial edge to each block. Figure 3.3 shows the result of this for the example in Fig. 3.2.

Now let us assume that we have a block that consists of two nodes and p parallel edges, $p \geq 4$. The following procedure is applied recursively. We separate into two blocks, the first with $\lfloor p/2 \rfloor$ edges and the second with the remainder. We add one artificial edge to each block. We can prove by induction that this procedure creates less than $2p$ new

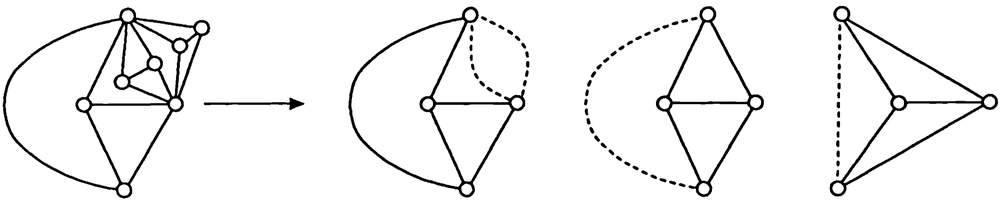


FIG. 3.2

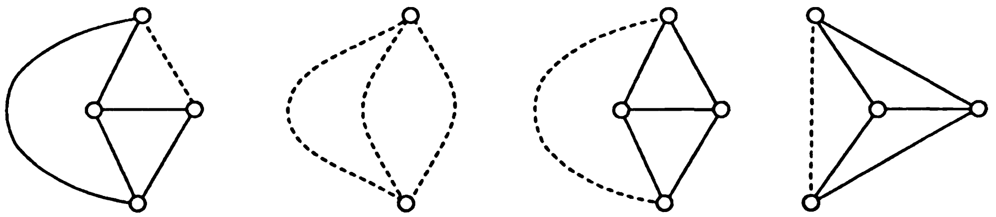


FIG. 3.3

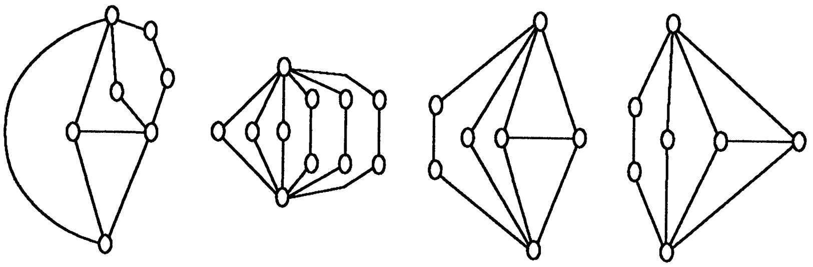


FIG. 3.4

artificial edges. Therefore the total number of edges is $O(n)$. Now the pieces are single edges, triangles, copies of K_4, and sets of at most three parallel edges. The first three types do not have parallel edges. Finally, Operation $\mathcal{O}$, given below, is applied to every artificial edge.

Operation $\mathcal{O}(uv)$. Remove the edge uv. Add the nodes w_1, w_2, and w_3; add the edges uw_1, vw_1, uw_2, w_2w_3, and w_3v.

Figure 3.4 shows the result of applying this to the pieces in Fig. 3.3.

Let us remark that the final pieces are series-parallel graphs with at most eleven nodes (like the second block in Fig. 3.4) and graphs obtained by applying $\mathcal{O}$ to K_4.

4. On the stable set polytope of graphs with no W_4 minor. In this section, we state our main result. The facets of $P(G)$ are not described in a simple way as "odd holes" or "cliques." We present a combinatorial procedure that produces all of them. We first present three theorems to derive "facets from facets." To make the notation less cumbersome, we use $a(u)$ instead of a_u to denote the coefficients of the inequalities.

THEOREM 4.1 (subdivision of an edge [15]). *Let $G = (V, E)$ be a graph and let uv be an edge of G and $\bar{G} = G\backslash uv$. Let $ax \leq \alpha$ be a facet-defining inequality of $P(G)$ different from $x(u) + x(v) \leq 1$. If $z = \max\ \{ax : x \in P(\bar{G})\}$ has a solution with $x(u) = x(v) = 1$, then $ax + \beta x(w) + \beta x(y) \leq z$ defines a facet of $P(G')$, where G' is the graph obtained from G by replacing the edge uv by the path (u, w, y, v), and $\beta = z - \alpha$.*

THEOREM 4.2 (contraction of an odd path [1]). *Let $G = (V, E)$ be a graph and let $ax \leq \alpha$ be a facet-defining inequality of $P(G)$. Suppose that G contains a path (pu, uv, vq) such that u and v are of degree 2. Assume also that $a(p) = a(u) = a(v) = \beta$. Let $G' = (V', E')$ be the graph obtained from G by replacing that path by the edge pq. Let*

$$\bar{a}(u) = a(u) \quad \textit{for } u \in V',$$

$$\bar{\alpha} = \alpha - \beta;$$

then $\bar{a}x \leq \bar{\alpha}$ defines a facet of $P(G')$.

THEOREM 4.3 (subdivision of a star [1]). *Let G be a graph and let $ax \leq \alpha$ be a nontrivial facet that is not an edge inequality. Let v be a node of G and let $N = \{v_0, \dots, v_{k-1}\}$ be its neighbor set. Suppose that, for each v_i, there is a stable set S_i such that $ax^{S_i} = \alpha$ and $S_i \cap N = \{v_i, v_{i+1}, \dots, v_{i+p-1}\}$, where $p \geq 1$ is a fixed integer and the indices are numbers modulo k. Suppose also that p and k are relatively prime and $a(v_0) = \cdots = a(v_k) = a(v)/p$. Let $G' = (V', E')$ be the graph obtained from G by adding on each edge vv_i a new node v'_i for $0 \leq i \leq k-1$. Set*

$$\bar{a}(u) = a(u) \quad \text{for } u \in V\backslash\{v\},$$

$$\bar{a}(v) = a(v)(k-p)/p,$$

$$\bar{a}(v'_i) = a(v)/p \quad \text{for } 0 \leq i \leq k-1,$$

$$\bar{\alpha} = \alpha + a(v)(k-p)/p.$$

Then $\bar{a}x \leq \bar{\alpha}$ defines a facet of $P(G')$.

It follows from Lemma 2.3 that, if we apply Operation $\mathcal{O}$ of §3 to K_4, the only nontrivial facets with support different from odd holes and edges, have as support a subdivision of K_4. Now we characterize those facets.

THEOREM 4.4. *If G is a subdivision of K_4, then the nontrivial facets of $P(G)$ are either odd holes or edges or have been obtained by applying Theorems* 4.1 *and* 4.3, *starting from the clique inequality of K_4.*

The proof of this is the subject of the next two sections. We prove that there are 16 cases to study and we give an explicit list of the facets for each case. We call them K_4 inequalities.

Now let G be a graph with n nodes that has no W_4 minor. Suppose that it is decomposed, as described in §3. For each piece, we have an explicit list of the facets; the total number of them is $O(n)$. The facets of $P(G)$ are obtained by composing those inequalities according to Theorem 2.5.

Our main result can be stated as follows.

THEOREM 4.5. *If G is a graph with no W_4 minor, then the nontrivial facets defining inequalities of $P(G)$ are either edge inequalities or can be constructed by composing inequalities from the following two families:* (i) *odd hole inequalities and* (ii) *a set of* 19 *K_4 inequalities.*

This last theorem gives a system that may have exponentially many inequalities. Suppose now that we use Theorem 2.4 instead of Theorem 2.5; i.e., we do not project the extra variables associated with the extra nodes. Then we can describe a polytope $Q = \{(x, y) : Ax + By \leq b\}$ such that $P(G) = \{x : \text{there is a vector } y, \text{ with } (x, y) \in Q\}$; i.e., $P(G)$ is a projection of Q. The decomposition of §3 gives a set of pieces that are series-parallel graphs with at most 11 nodes and copies of K_4 with some edges replaced by a 5-cycle. The total number of nodes is $O(n)$. Thus the number of variables and the number of inequalities in the system that defines Q is $O(n)$. Moreover, the coefficients in those inequalities are integer numbers of absolute value at most 2.

Therefore, for this class of graphs, the stable set problem can be formulated as a linear program of polynomial size.

For a polytope P the so-called *separation* problem is: Given a vector $\bar{x}$, decide whether $\bar{x} \in P$ and, if not, find a hyperplane that separates $\bar{x}$ from P.

Grötschel, Lovász, and Schrijver [8], [9] have shown that, if the optimization problem can be solved in polynomial time, then the separation problem can also be solved in polynomial time by means of the ellipsoid method.

In our case, by the Farkas lemma, if $\bar{x} \notin P(G)$, there is a vector π such that

$$\pi B = 0, \quad \pi \geq 0, \quad \text{and} \quad \pi(b - A\bar{x}) < 0.$$

So $\pi Ax \leq \pi b$ is the required inequality. Thus, the separation problem can be solved in polynomial time by means of any polynomial algorithm for linear programming; cf. Khachiyan [12] and Karmarkar [11], for instance.

Solving the stable set problem with a cutting plane approach based on this separation algorithm is equivalent to applying Benders decomposition to the linear program

$$\text{maximize } wx \quad \text{s.t. } Ax + By \leq b.$$

5. Technical lemmas. In this section, we present a series of lemmas that lead to the characterization of $P(G)$, when G is a subdivision of K_4. First, note the following remarks.

Remark 5.1. It follows from the results of §2 that it is enough to characterize the polytope for graphs that are obtained by replacing some edges of K_4 by paths with two or three edges.

Remark 5.2. It is enough to consider the case where the four faces of the graph are odd (a planar graph has an even number of odd faces). If only two of them are odd, then there is a node that covers them; i.e., the removal of this node will leave a bipartite graph. In this case, the graph is t-perfect, as shown by Fonlupt and Uhry [5].

Remark 5.3. Suppose that we have a graph that has been obtained from K_4 by replacing some edges by paths of two edges, with the additional condition that every original node has at least one incident edge that has not been replaced. Since we should have four odd faces, the only graph of this kind to be studied is the graph of Fig. 5.1. This has been shown to be t-perfect by Gerards and Schrijver [7].

Remark 5.4. Consider a graph $G = (V, E)$ that is a subdivision of K_4 and let $ax \leq \alpha$ with $\alpha > 0$ be a facet of $P(G)$ whose support is not a clique or an odd hole. Since $G\backslash v$ is a series-parallel graph, we have that $a(v) > 0$ for all $v \in V$.

From the first three remarks, it follows that we must study the 16 cases shown in Fig. 5.2. We first state a lemma that will be used in this section.

LEMMA 5.5. *Let $ax \leq \alpha$ be a facet of $P(G)$. Let u be a node of degree 2 in G_a and let v, w be the neighbors of u in G_a. Then $a(v) \geq a(u) \leq a(w)$.*

Proof. The equation

$$ax = \alpha \tag{5.1}$$

defines a hyperplane different from those of $x(v) + x(w) = 2$ and $x(v) - x(w) = 0$. So there is a stable set S such that $ax^S = \alpha$ and $|S \cap \{v, w\}| = 1$. Assume that $v \in S$. Since $(S\backslash\{v\}) \cup \{u\}$ is a stable set, we can conclude that $a(u) \leq a(v)$.

Since (5.1) also defines a hyperplane different from that of $x(u) + x(v) = 1$, there is a stable set T such that $ax^T = \alpha$, and $T \cap \{u, v, w\} = \{w\}$. This implies that $a(u) \leq a(w)$. □

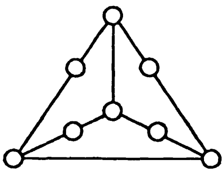


FIG. 5.1

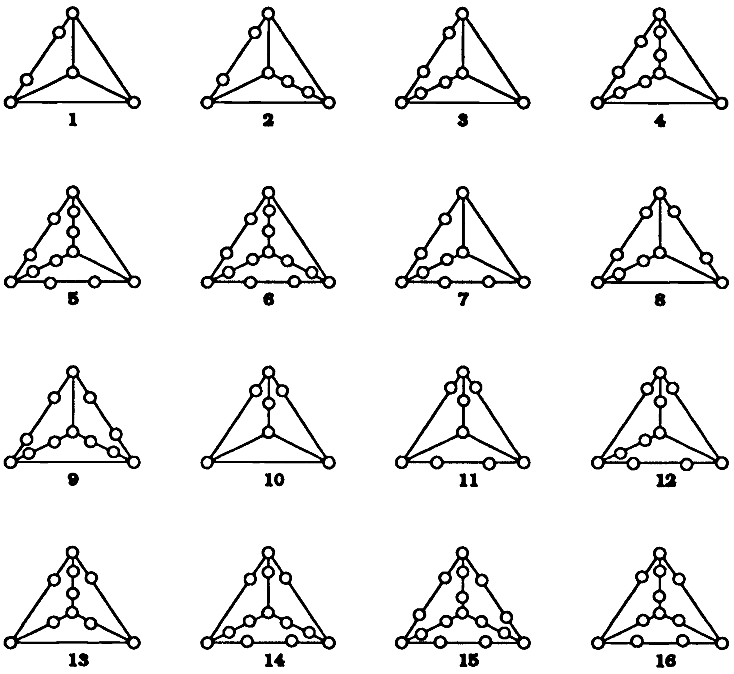


FIG. 5.2

The following lemma will be used to solve nine cases.

LEMMA 5.6. *Let $G = (V, E)$ be a graph and let (u, w, y, v) be a path in G, where the nodes w and y have degree 2. Let G' be the graph obtained from G by replacing this path by one edge; see Fig.* 5.3. *Let $N(u)$ and $N(v)$ be the neighbor sets of u and v in G, respectively. Suppose that $ax \leq \alpha$ is the only facet-defining inequality of $P(G')$ whose support is G' and let $bx \leq \beta$ be the facet of $P(G)$ obtained from $ax \leq \alpha$ by the procedure described in Theorem* 4.1. *Suppose that, for every facet $\bar{a}x \leq \bar{\alpha}$ of $P(G)$ whose support is G, there exists a stable set S of G such that $\bar{a}x^S = \bar{\alpha}$ and either $S \cap (N(u)\backslash\{w\}) = \emptyset$ or $S \cap (N(v)\backslash\{y\}) = \emptyset$. Then $bx \leq \beta$ is the only facet of $P(G)$ whose support is G.*

Proof. Let $\bar{a}x \leq \alpha$ be a facet of $P(G)$, whose support is G. Lemma 5.5 implies that $\bar{a}(w) = \bar{a}(y) \leq \min\{\bar{a}(u), \bar{a}(v)\}$. If there exists a stable set S of G such that $\bar{a}x^S = \bar{\alpha}$ and $S \cap (N(u)\backslash\{w\}) = \emptyset$, say, then $w \in S$ and $S' = (S\backslash\{w\}) \cup \{u\}$ is also a stable set in G. Hence $\bar{a}(u) \leq \bar{a}(w) = \bar{a}(y) \leq \bar{a}(u)$. From Theorem 4.2, we have that the inequality

$$a'x \leq \alpha' \tag{5.2}$$

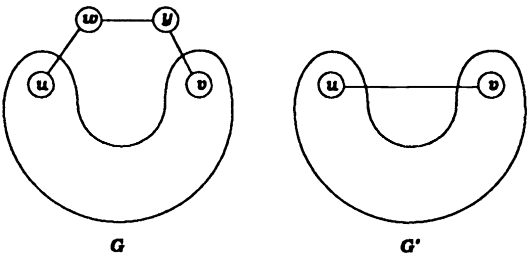


FIG. 5.3

defines a facet of $P(G')$, where

$$a'(i) = \bar{a}(i) \quad \text{for } i \in V\backslash\{w, y\}$$

and

$$\alpha' = \bar{\alpha} - \bar{a}(u).$$

Since the support of (5.2) is G', we have that $a' = a$ and $\alpha' = \alpha$. Thus $\bar{a} = b$ and $\bar{\alpha} = \beta$. The proof is complete. □

The next lemma will allow us to solve two cases.

LEMMA 5.7. *Let G be a graph obtained from K_4 in such a way that at least two edges have not been subdivided. Let G' be the graph obtained from G by replacing one of those two edges by a path of three edges; see Fig.* 5.4. *If $ax \le \alpha$ is the only facet of $P(G)$ having G as support, then the only facet of $P(G')$ having G' as support is the inequality $a'x \le \alpha'$ obtained by applying the procedure of Theorem* 4.1 *to $ax \le \alpha$.*

Proof. Let $bx \le \beta$ be a facet of $P(G')$ whose support is G'. Let C be the odd hole of G' defined by the edge $\{1, 3\}$ and the paths 1-2 and 2-3. Since $bx \le \beta$ is different from the facet associated with C, there is a stable set S such that $bx^S = \beta$ and $|C \cap S| < (|C| - 1)/2$.

If $|C| = 3$, then $S \cap C = \emptyset$. If $|C| = 5$ or $|C| = 7$, then S can be chosen so that $S \cap C = Z$, where Z is the set of nodes of C adjacent to the node 2 and different from nodes 1 and 3. In these three cases, we have that $S \cap (N(1)\backslash\{w\}) = \emptyset$. From the previous lemma we have that $b = \rho a'$ and $\beta = \rho\alpha'$, for some $\rho > 0$. □

The next three lemmas will enable us to solve four cases.

LEMMA 5.8. *Let v be a node of G. Let $N(v)$ be the neighbor set of v. Let $ax \le \alpha$ be a facet-defining inequality of $P(G)$ whose support is not an edge; then*

$$a(v) \le a(N(v)\backslash\{u\})$$

for all $u \in N(v)$.

Proof. Let $u \in N(v)$. Since $ax \le \alpha$ is a facet not associated with an edge, then there is a stable set S such that

$$ax^S = \alpha$$

and $\{u, v\} \cap S = \emptyset$.

Then $(S\backslash(N(v)\backslash\{u\})) \cup \{v\}$ is a stable set in G. This implies that

$$a(v) \le a(N(v)\backslash\{u\}). \quad \square$$

LEMMA 5.9. *Let G be a graph as in Fig.* 5.5, *where the dashed lines represent paths with one or more edges, and v (respectively, w) is a node adjacent to v_3 (respectively, v_1) in the path that replaces the edge v_3v_4 (respectively, v_1v_4) of K_4. If $ax \le \alpha$ defines a facet*

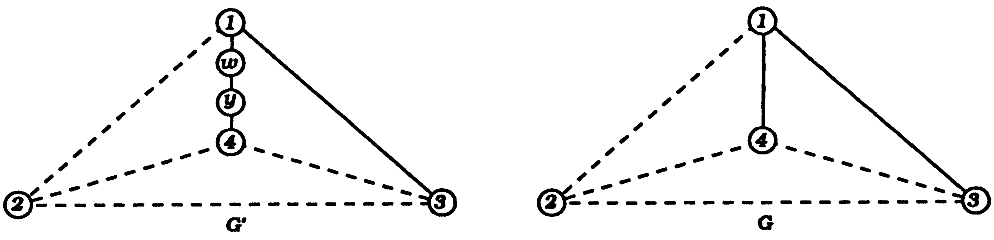


FIG. 5.4

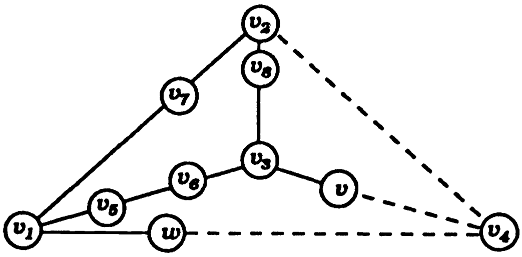


FIG. 5.5

of $P(G)$ *whose support is* G, *then we have the following*:

(a) *Either* (i) $a(v_1) = a(v_5) + a(v_7)$ *and* $a(v_3) = a(v_6) + a(v_8)$, *or* (ii) $a(v_2) = a(v_7) + a(v_8)$, $a(v_1) = a(w)$, *and* $a(v_3) = a(v)$;

(b) *If the edge* v_2v_4 *is subdivided and* u *is the node adjacent to* v_2 *in this path, then*

(b1) *Either* (i) *and* $a(v_2) = a(u)$ *hold, or* (ii) *holds*,

(b2) *If the path between* u *and* v_4 *is* (uy, yv_4) *and* $a(v_2) = a(u)$, *then* $ax \leq \alpha$ *is obtained from a facet of* $P(G')$ *using the procedure of Theorem* 4.1, *where* G' *is the graph obtained from* G *by contracting the edges* uy *and* yv_4.

Proof. (a) Let C denote the cycle $(v_1, v_5, v_6, v_3, v_8, v_2, v_7, v_1)$. Since $ax \leq \alpha$ has as support the graph G and C is an odd hole, there is a stable set S in G such that $|S \cap C| < 3$ and $ax^S = \alpha$.

Case 1. $\{v_1, v_3\} \subseteq S$.

Thus $S \cap C = \{v_1, v_3\}$. Since $(S \backslash \{v_1\}) \cup \{v_5, v_7\}$ is a stable set, it follows that $a(v_1) \geq a(v_5) + a(v_7)$. From Lemma 5.8, we have that $a(v_1) = a(v_5) + a(v_7)$.

Since $(S \backslash \{v_3\}) \cup \{v_6, v_8\}$ is also a stable set in G, we obtain $a(v_6) + a(v_8) = a(v_3)$ in a similar way.

Case 2. $\{v_1, v_3\} \not\subseteq S$.

We should have that $\{v_1, v_3\} \cap S = \emptyset$. If, for instance, $v_1 \in S$ and $v_3 \notin S$, then $\{v_2, v_6\} \subseteq S$ or $\{v_8, v_6\} \subseteq S$, and $|S \cap C| = 3$, which is a contradiction.

Therefore S must contain v_2 and a node from $\{v_5, v_6\}$, say v_6. Then $(S \backslash \{v_2\}) \cup \{v_7, v_8\}$ is a stable set, which implies that $a(v_2) \geq a(v_7) + a(v_8)$, and, from Lemma 5.8, we have that $a(v_2) = a(v_7) + a(v_8)$.

Furthermore, $w \in S$; otherwise, $v_1 \in S$, which is a contradiction. So $(S \backslash \{w\}) \cup \{v_1\}$ is a stable set. Then $a(v_1) \leq a(w)$, and, from Lemma 5.5, we have that $a(v_1) = a(w)$. Also, we have $v \in S$. If not, $(S \backslash \{v_6\}) \cup \{v_3, v_5\}$ is a stable set. Since $a(v_5) = a(v_6)$ and $a(v_3) > 0$, we have a contradiction.

Since $(S \backslash \{v, v_6\}) \cup \{v_5, v_3\}$ is a stable set, we have that $a(v_3) \leq a(v)$, and hence $a(v_3) = a(v)$.

We now prove (b).

(b1) If $\{v_1, v_3\} \subseteq S$, then (i) holds. Since $S \cap C = \{v_1, v_3\}$, we should have that $u \in S$; otherwise, $v_2 \in S$, which is a contradiction. Since $(S \backslash \{u\}) \cup \{v_2\}$ is a stable set, we have that $a(v_2) \leq a(u)$ and, from Lemma 5.5, we can deduce that $a(v_2) = a(u)$.

(b2) Since $a(v_2) = a(u) = a(y)$, the statement follows from Theorem 4.2. □

LEMMA 5.10. *Let* G *be a graph as in Fig.* 5.6, *where the dashed lines represent paths with one or two edges and* u (*respectively*, v, w) *is a node adjacent to* v_1 (*respectively*, v_2, v_3) *in the path that replaces the edge* v_1v_4 (*respectively*, v_2v_4, v_3v_4) *of* K_4. *If* $ax \leq \alpha$ *defines a facet of* $P(G)$ *whose support is* G, *then either*

(i) $a(v_1) = a(v_7) + a(v_8)$, $a(v_2) = a(v_5) + a(v_6)$, *and* $a(v_3) = a(v_{10}) + a(v_9)$, *or*

(ii) $a(v_1) = a(u)$, $a(v_2) = a(v)$, *and* $a(v_3) = a(w)$.

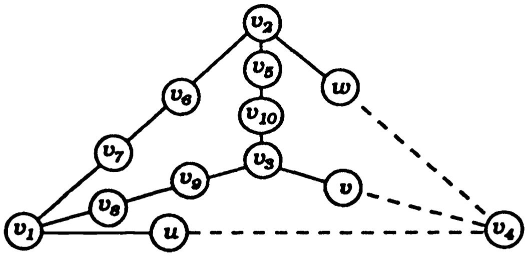


FIG. 5.6

Proof. Let C denote the cycle $(v_1, v_8, v_9, v_3, v_{10}, v_5, v_2, v_6, v_7, v_1)$. Since $ax \leq \alpha$ has as support the graph G and since C is an odd hole, there is a stable set S in G such that $|S \cap C| < 4$ and $ax^S = \alpha$.

Case 1. $S \cap C = \{v_1, v_2, v_3\}$.

Let

$$S_1 = (S \backslash \{v_1\}) \cup \{v_7, v_8\},$$
$$S_2 = (S \backslash \{v_2\}) \cup \{v_5, v_6\},$$
$$S_3 = (S \backslash \{v_3\}) \cup \{v_{10}, v_9\}.$$

Clearly, S_1, S_2, and S_3 are stable sets in G. Thus we should have

$$a(v_1) \geq a(v_7) + a(v_8),$$
$$a(v_2) \geq a(v_5) + a(v_6),$$
$$a(v_3) \geq a(v_{10}) + a(v_9).$$

From Lemma 5.8, it follows that these inequalities should be equations.

Case 2. $\{v_1, v_2, v_3\} \not\subseteq S$.

CLAIM 1. *We have that* $\{v_1, v_2, v_3\} \cap S = \emptyset$.

Let us assume, for instance, that $v_1 \in S$. If $\{v_2, v_3\} \cap S = \emptyset$, then $\{v_6, v_9\} \subseteq C$ and $|\{v_{10}, v_5\} \cap S| = 1$; hence $|S \cap C| = 4$, a contradiction.

Suppose that $\{v_2, v_3\} \cap S \neq \emptyset$ and let us assume, for instance, that $\{v_2, v_3\} \cap S = \{v_2\}$; then we should have that $\{v_{10}, v_9\} \subseteq S$ and $|S \cap C| = 4$, a contradiction, which proves Claim 1.

CLAIM 2. *We have that* $\{u, v, w\} \subseteq S$.

Assume, for instance, that $u \notin S$. Since we can assume that $S \cap C = \{v_6, v_9, v_5\}$, we would have that $v_1 \in S$, which is a contradiction. This proves Claim 2.

Let

$$S_1 = (S \backslash \{u, v_7, v_8\}) \cup \{v_1, v_6, v_9\},$$
$$S_2 = (S \backslash \{v, v_{10}, v_9\}) \cup \{v_3, v_5, v_8\},$$
$$S_3 = (S \backslash \{w, v_5, v_6\}) \cup \{v_2, v_{10}, v_7\}.$$

These node sets define stable sets of G. Since $a(v_7) + a(v_8) = a(v_6) + a(v_9)$, we have that $a(v_1) \leq a(u)$. It follows from Lemma 5.5 that $a(v_1) = a(u)$.

In a similar way, we can prove that $a(v_2) = a(w)$ and $a(v_3) = a(v)$. □

6. The stable set polytope of a subdivision of K_4. The purpose of this section is to prove Theorem 4.4.

First, we can apply Lemma 5.6 to prove our claim for the graphs of Cases 1–9 in Fig. 5.2. In each case, there is a unique facet whose support is the whole graph, obtained by applying Theorem 4.1 starting from the clique inequality of K_4. Now consider the graph of Case 10. It has seven nodes and seven maximal stable sets. Then there is a unique facet that may have this graph as support; this is the facet obtained by applying Theorem 4.3 to the clique inequality of K_4. Starting from this graph, we can apply Lemma 5.7 to prove our claim for the graphs in Cases 11 and 12. Now we must study the graphs of Cases 13–16. In all these cases, we denote by $ax \leq \alpha$ a facet whose support is the whole graph. We begin with the graph of Case 13; see Fig. 6.1.

Lemma 5.9 implies that either (i) $a(v_2) = a(v_5) + a(v_6)$ and $a(v_3) = a(v_8) + a(v_9)$, or (ii) $a(v_1) = a(v_5) + a(v_9)$ and $a(v_2) = a(v_7)$. It also implies that either (iii) $a(v_2) = a(v_6) + a(v_7)$ and $a(v_3) = a(v_8) + a(v_{10})$, or (iv) $a(v_4) = a(v_7) + a(v_{10})$, $a(v_2) = a(v_5)$, and $a(v_3) = a(v_9)$. Thus we have that either (i) and (iii) hold or (ii) and (iv) hold.

Consider the cycle $C = (v_1, v_4, v_7, v_2, v_5, v_1)$. Since $ax \leq \alpha$ has as support the graph G and C is an odd hole, there is a stable set S in G such that $|S \cap C| < 2$ and $ax^S = \alpha$. Then $S \cap C = \{v_2\}$; therefore $S = \{v_2, v_8, v_9, v_{10}\}$. Since $(S\backslash\{v_9\}) \cup \{v_1\}$ and $(S\backslash\{v_{10}\}) \cup \{v_4\}$ are stable sets, we have that $a(v_1) \leq a(v_9)$ and $a(v_4) \leq a(v_{10})$. Lemma 5.5 implies that (v) $a(v_1) = a(v_9)$ and (vi) $a(v_4) = a(v_{10})$.

In the same way, by considering the cycle $C = (v_3, v_9, v_1, v_4, v_{10}, v_3)$, we can prove that (vii) $a(v_1) = a(v_5)$ and (viii) $a(v_4) = a(v_7)$.

Consider (i), (iii), and (v)–(viii). This implies that

$$a(v_1) = a(v_4) = a(v_5) = a(v_7) = a(v_9) = a(v_{10})$$

and

$$a(v_2) = a(v_3) = a(v_6) + a(v_7).$$

Since the inequality $x(v_2) + x(v_6) + x(v_8) + x(v_3) \leq 2$ is valid for $P(G)$, there is a stable set T such that $ax^T = \alpha$ and x^T satisfies $x(v_2) + x(v_6) + x(v_8) + x(v_3) < 2$. We can choose $T = \{v_5, v_7, v_8, v_9, v_{10}\}$.

Consider the set S defined above. Since $ax^S = ax^T$, we have that $a(v_5) = a(v_8) = a(v_6)$. Therefore $ax \leq \alpha$ represents the inequality

$$2x(v_2) + 2x(v_3) + x(v_1) + \sum_{j=4}^{10} x(v_j) \leq 5.$$

Condition (v) implies that (ii) and (iv) cannot hold.

Now we study the graph of Case 14; see Fig. 6.2.

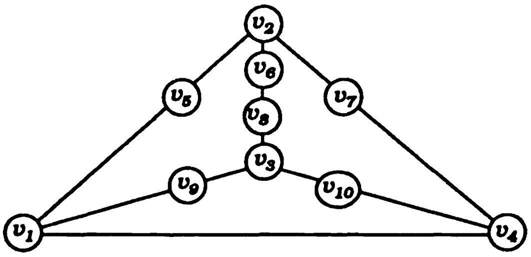


FIG. 6.1

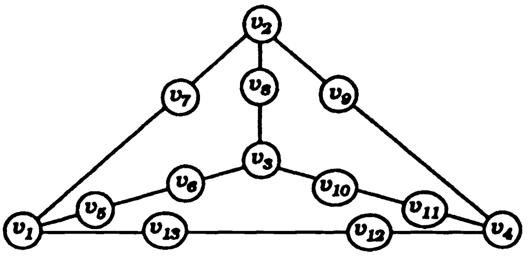


FIG. 6.2

Lemma 5.9 implies that either (i) $a(v_1) = a(v_5) + a(v_7)$, $a(v_3) = a(v_6) + a(v_8)$, and $a(v_2) = a(v_9)$, or (ii) $a(v_2) = a(v_7) + a(v_8)$, $a(v_3) = a(v_{10})$, and $a(v_1) = a(v_{13})$. It also implies that either (iii) $a(v_4) = a(v_{11}) + a(v_9)$, $a(v_3) = a(v_8) + a(v_{10})$, and $a(v_2) = a(v_7)$, or (iv) $a(v_2) = a(v_8) + a(v_9)$, $a(v_3) = a(v_6)$, and $a(v_4) = a(v_{12})$, and that either (v) $a(v_4) = a(v_{12}) + a(v_9)$, $a(v_1) = a(v_7) + a(v_{13})$, and $a(v_2) = a(v_8)$, or (vi) $a(v_2) = a(v_7) + a(v_9)$, $a(v_1) = a(v_5)$, and $a(v_4) = a(v_{11})$.

We have that either (i), (iii), and (v) hold or (ii), (iv), and (vi) hold. Consider (i), (iii), and (v). This implies that

$$\beta = a(v_{10}) = a(v_6) = a(v_{11}) = a(v_5) = a(v_{12}) = a(v_{13}),$$
$$\gamma = a(v_2) = a(v_7) = a(v_8) = a(v_9),$$
$$\beta + \gamma = a(v_1) = a(v_4) = a(v_3).$$

Lemma 5.8 implies that $\beta \geq \gamma$. Consider the cycle $C = (v_1, v_5, v_6, v_3, v_{10}, v_{11}, v_4, v_{12}, v_{13}, v_1)$. Since $ax \leq \alpha$ has as support the graph G and C is an odd hole, there is a stable set S in G such that $|S \cap C| < 4$ and $ax^S = \alpha$. Then $S = \{v_1, v_2, v_3, v_4\}$. Since $(S \backslash \{v_4\}) \cup \{v_{11}, v_{12}\}$ is a stable set, it follows that $\beta \leq \gamma$. Therefore we have the inequality

$$2x(v_1) + 2x(v_3) + 2x(v_4) + x(v_2) + \sum_{j=5}^{13} x(v_j) \leq 7.$$

Now consider (ii), (iv), and (vi). This implies that

$$\beta = a(v_1) = a(v_5) = a(v_6) = a(v_3) = a(v_{10}) = a(v_{11}) = a(v_4) = a(v_{12}) = a(v_{13}),$$
$$\gamma = a(v_7) = a(v_8) = a(v_9),$$
$$2\gamma = a(v_2).$$

Lemma 5.5 implies that $\beta \geq \gamma$. Consider the cycle $C = (v_1, v_5, v_6, v_3, v_{10}, v_{11}, v_4, v_{12}, v_{13}, v_1)$. Since $ax \leq \alpha$ has as support the graph G and C is an odd hole, there is a stable set S in G such that $|S \cap C| < 4$ and $ax^S = \alpha$. Then $S = \{v_5, v_7, v_8, v_9, v_{10}, v_{13}\}$. Since $(S \backslash \{v_9\}) \cup \{v_4\}$ is a stable set, it follows that $\beta \leq \gamma$. Therefore we have the inequality

$$2x(v_2) + \sum_{j \neq 2} x(v_j) \leq 6.$$

Now we study Case 15; see Fig. 6.3.

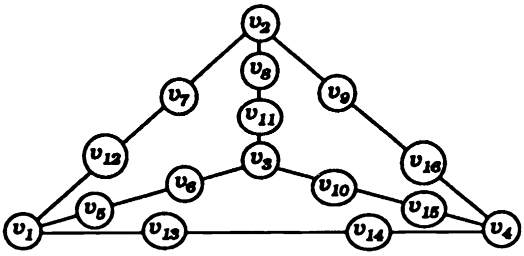


FIG. 6.3

Lemma 5.10 implies that either (i) $\beta = a(v_j)$ for $5 \leq j \leq 16$ and $2\beta = a(v_j)$ for $1 \leq j \leq 4$ or (ii) $\beta = a(v_j)$, for $1 \leq j \leq 16$. In the first case, we have the inequality

$$2 \sum_{j=1}^{4} x(v_j) + \sum_{j=5}^{16} x(v_j) \leq 9.$$

In the second case, we have the inequality

$$\sum x(v_j) \leq 7.$$

Now we study Case 16; see Fig. 6.4.

Part (b1) of Lemma 5.9 implies that either (i) $a(v_2) = a(v_7) + a(v_8)$, $a(v_3) = a(v_6) + a(v_{11})$, and $a(v_1) = a(v_5)$ or (ii) $a(v_1) = a(v_6) + a(v_7)$, $a(v_3) = a(v_{10})$, and $a(v_2) = a(v_9)$. It also implies that either (iii) $a(v_1) = a(v_5) + a(v_7)$, $a(v_4) = a(v_9) + a(v_{12})$, and $a(v_2) = a(v_8)$ or (iv) $a(v_2) = a(v_7) + a(v_9)$, $a(v_1) = a(v_6)$, and $a(v_4) = a(v_{10})$. We have that either (i) and (iv) hold or (ii) and (vii) hold.

Consider (i) and (iv). This implies that $a(v_1) = a(v_5)$. It follows from Lemma 5.11 (b2) that the inequality $ax \leq \alpha$ is obtained from a facet of $P(G')$, using the procedure of Theorem 4.1, where G' is obtained from G by contracting the edges v_5v_{12} and $v_{12}v_4$. The graph G' is that of Case 13, which has already been studied. For this, $P(G')$ has only one facet whose support is G', which is the inequality

$$2x(v_2) + 2x(v_3) + x(v_1) + \sum_{j=4}^{10} x(v_j) \leq 5.$$

Then $ax \leq \alpha$ should be

$$x(v_1) + 2x(v_2) + 2x(v_3) + \sum_{j=4}^{12} x(v_j) \leq 6.$$

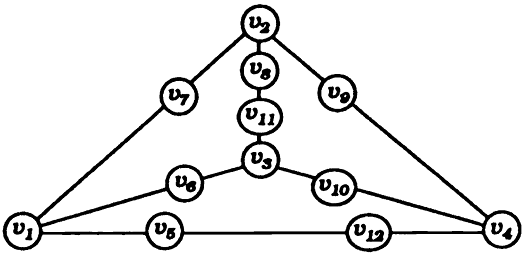


FIG. 6.4

If (ii) and (iii) hold, then $a(v_2) = a(v_8)$. In a similar way, we can prove that $ax \leq \alpha$ is

$$2x(v_1) + x(v_2) + x(v_3) + 2x(v_4) + \sum_{j=5}^{12} x(v_j) \leq 6.$$

It is easy to see that all the inequalities derived in this section can be obtained by applying Theorems 4.1 and 4.3, starting from the clique inequality of K_4. This completes the proof of Theorem 4.4.

The 18 inequalities derived in this section, together with the clique inequality of K_4, form a family of 19 K_4 inequalities, referred to in Theorem 4.5.

7. Some examples. In this section, we apply the combinatorial procedure that describes the facets of $P(G)$ for graphs with no W_4 minor. Consider the graphs of Figs. 7.1(a) and 7.1(b).

The constraint

$$2x(u_1) + x(u_2) + 2x(u_3) + 2x(u_4) + \sum_{j \geq 5} x(u_j) \leq 7$$

defines a facet for the polytope of the first graph, and the constraint

$$x(v_1) + 2x(v_2) + 2x(v_3) + \sum_{j \geq 4} x(v_j) \leq 5$$

defines a facet for the second one.

By identifying the nodes $\{u_1, v_1\}$ and $\{u_4, v_2\}$ and deleting a 5-cycle, we obtain the graph of Fig. 7.2.

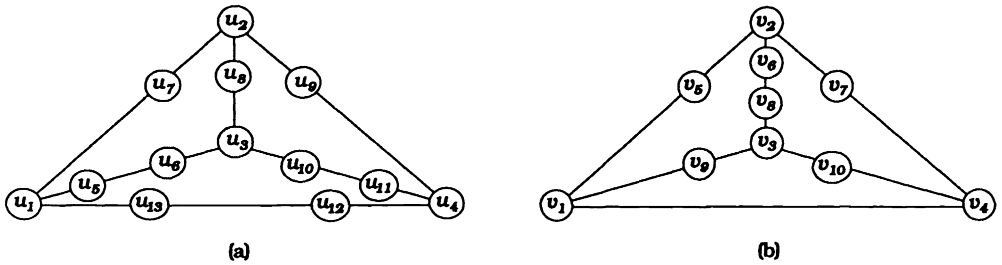


FIG. 7.1

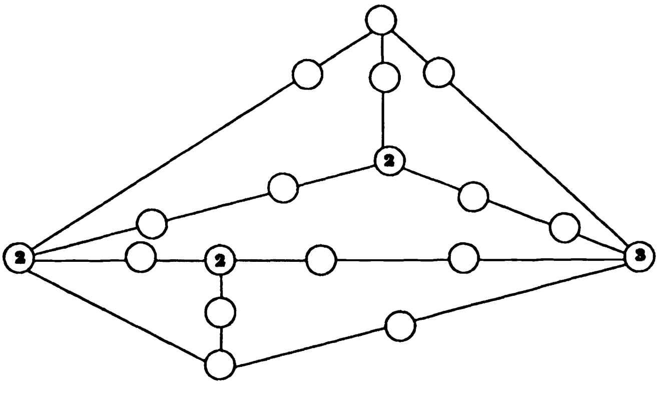


FIG. 7.2

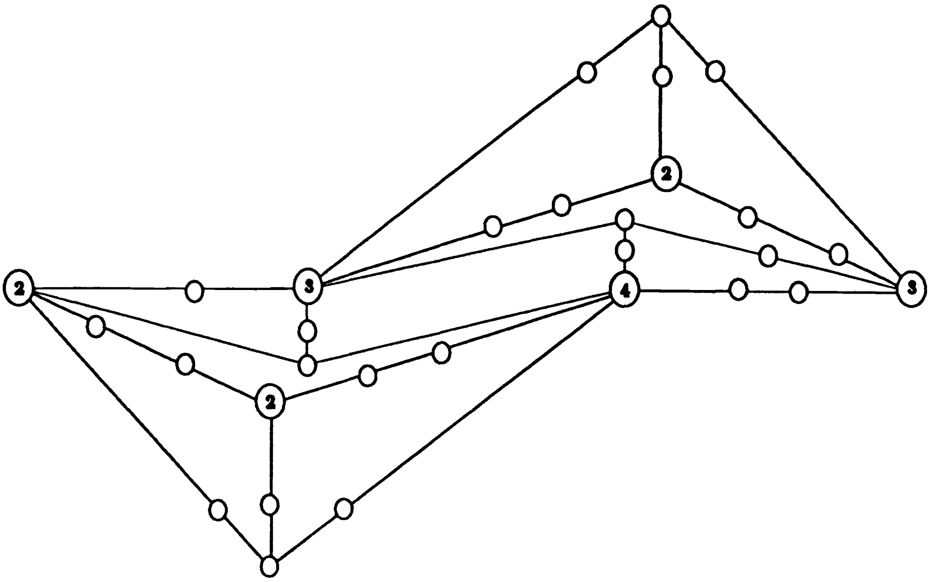


FIG. 7.3

Theorem 2.5 gives a facet-defining inequality, whose support is the whole graph, and its coefficients different from 1 appear in the figure. The value of the right-hand side is 10. Here v_5 plays the role of w_1, and u_{12} and u_{13} play the roles of w_2 and w_3, respectively.

Now consider the graph of Fig. 7.3. It has been obtained by composing the graph of Fig. 7.2 with itself. Again, Theorem 2.5 shows that there is a facet-defining inequality whose support is the whole graph and whose coefficients different from 1 are in the figure. We can state the following.

Remark 7.1. Given any integer $p > 0$, there exists a graph G with no W_4 minor, such that $P(G)$ has a facet with coefficients $1, 2, \dots, p$.

Acknowledgments. We thank the referees for their suggestions regarding the presentation.

REFERENCES

[1] F. BARAHONA AND A. R. MAHJOUB, *Compositions of graphs and polyhedra* II: *stable sets*, Research Report CORR 87-47, University of Waterloo, Canada, 1987; SIAM J. Discrete Math., 7 (1994), pp. 359–371, this issue.

[2] M. BOULALA AND J. P. UHRY, *Polytope des indépendants d'un graphe série-parallele*, Discrete Math., 27 (1979), pp. 225–243.

[3] V. CHVÁTAL, *On certain polytopes associated with graphs*, J. Combin. Theory, Ser. B, 18 (1975), pp. 138–154.

[4] J. EDMONDS, *Maximum matching and a polyhedron with* (0, 1)-*vertices*, J. Res. National Bureau of Standards, 69B (1965), pp. 125–130.

[5] J. FONLUPT AND J. P. UHRY, *Transformations which preserve perfectness and h-perfectness of graphs*, Ann. Discrete Math., 16 (1985), pp. 83–95.

[6] H. GAN AND E. L. JOHNSON, *Four problems on graphs with excluded minors*, Math. Programming, 45 (1989), pp. 311–330.

[7] A. M. H. GERARDS AND A. SCHRIJVER, *Matrices with the Edmonds–Johnson property*, Combinatorica, 6 (1986), pp. 365–379.

[8] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–191.

[9] ———, *Geometric Algorithms and Combinatorial Optimization*, Springer, Berlin, New York, 1988.

[10] R. Halin, *Graphentheorie* II, Wissenschaftliche Buchgesellschaft, Darmstadt, 1981.

[11] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[12] L. Khachiyan, *A polynomial algorithm in linear programming*, Soviet Math. Dokl., 20 (1979), pp. 191–194.

[13] A. R. Mahjoub, *On the stable set polytope of a series-parallel graph*, Math. Programming, 40 (1988), pp. 53–57.

[14] M. Padberg, *On the facial structure of set packing problems*, Math. Programming, 5 (1973), pp. 199–215.

[15] L. Wolsey, *Further facet generating procedures for vertex packing polytopes*, Math. Programming, 11 (1976), pp. 158–163.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 390–402, August 1994

005

COMPOSITIONS OF GRAPHS AND POLYHEDRA IV: ACYCLIC SPANNING SUBGRAPHS*

FRANCISCO BARAHONA†, JEAN FONLUPT‡, AND ALI RIDHA MAHJOUB§

Abstract. Given a directed graph D that has a two-vertex cut, this paper describes a technique to derive a linear system that defines the acyclic subgraph polytope of D from systems related to the pieces. It also gives a technique to describe facets of this polytope by composition of facets for the pieces. The authors prove that, if the systems for the pieces are totally dual integral (TDI), then the system for D is also. The authors prove that the "cycle inequalities" form a TDI system for any orientation of K_5. These results are combined with Lucchesi–Younger theorem and a theorem of Wagner to prove that, for graphs with no $K_{3,3}$ minor, the cycle inequalities characterize the acyclic subgraph polytope and form a TDI system. This shows that, for this class of graphs, the cardinality of a minimum feedback set is equal to the maximum number of arc disjoint cycles. For planar graphs, this is a consequence of the Lucchesi–Younger theorem.

Key words. polyhedral combinatorics, compositions of polyhedra, acyclic subgraph polytope

AMS subject classifications. 05C85, 90C27

1. Introduction. Given a directed graph $D = (V, A)$, we say that $D' = (V, A')$ is a subgraph of D if $A' \subseteq A$. Given $S \subseteq A$, the incidence vector of S, $x^S \in \Re^A$ is defined by

$$x^S(i,j) = \begin{cases} 1 & \text{if } (i,j) \in S, \\ 0 & \text{if } (i,j) \in A \setminus S. \end{cases}$$

The *acyclic subgraph polytope* of D, denoted by $P(D)$, is the convex hull of incidence vectors of arc sets S such that $D_S = (V, S)$ has no directed cycle. Given a weight function $w : A \rightarrow \Re$, the problem of finding a maximum weighted cyclic subgraph can be formulated as the linear program

$$\text{maximize } wx \quad \text{s.t. } x \in P(D).$$

The polytope $P(D)$ is full-dimensional. This implies that (up to multiplication by a positive constant) there is a unique nonredundant inequality system $Ax \leq b$ such that $P(D) = \{x : Ax \leq b\}$. These inequalities define the *facets* of $P(D)$. The acyclic subgraph problem is NP-hard, so finding a complete characterization of $P(D)$ seems to be very difficult. On the other hand, Lucchesi and Younger [9] characterized $P(D)$ for planar graphs. Grötschel, Jünger, and Reinelt [6], [7] characterized several facet-defining inequalities of $P(D)$ and used them to design a cutting plane algorithm.

In this paper, we study directed graphs that have a two-vertex cutset. We show how to derive a system that defines $P(D)$ from systems associated with the pieces. This also gives a technique to derive new facets defining inequalities by composition of known

* Received by the editors June 4, 1990; accepted for publication (in revised form) May 20, 1993.

† IBM T. J. Watson Research Center, Yorktown Heights, New York 10598.

‡ Laboratoire ARTEMIS, Université Joseph Fourier, Grenoble, France.

§ Department of Statistics, King Saud University, Riyadh, Saudi Arabia. Present address, Laboratoire d'Informatique de Brest (LIBr), Université de Bretagne Occidentale, 6 Avenue Le Gorgeu, 29287 Brest Cedex, France.

facets. We also prove that, if the two systems are totally dual integral (TDI), then the new system is, also.

The most natural system of inequalities that we can think of is

$$\sum_{(i,j)\in C} x(i,j) \leq |C| - 1 \quad \text{for every directed cycle } C, \tag{1.1}$$

$$0 \leq x(i,j) \leq 1 \quad \text{for every arc } (i,j). \tag{1.2}$$

Inequalities (1.1) will be called *cycle inequalities*. Lucchesi and Younger [9] proved that, for planar graphs, (1.1), (1.2) is a TDI system and defines $P(D)$.

Wagner [10] proved that graphs with no $K_{3,3}$ minor can be decomposed into planar graphs and copies of K_5. We prove that, for any orientation of K_5, the system (1.1), (1.2) is TDI. We combine Wagner's theorem with the theorem of Lucchesi and Younger and our composition techniques to prove that, for graphs with no $K_{3,3}$ minor, the system (1.1), (1.2) is TDI and defines $P(D)$. This implies that, for this class of graphs, the cardinality of a minimum feedback set is equal to the maximum number of arc disjoint cycles.

The present paper should be considered as a revision of [3]; this type of composition was studied there, but the results on dual integrality are new.

If $G = (V, E)$ is an undirected graph, we say that G contains H as a minor if H can be obtained from G by a sequence of deletions and contractions of edges. An orientation of G is a directed graph that contains exactly one of the arcs (i, j) or (j, i) whenever $ij \in E$. The symmetric digraph $D(G) = (V, A)$ associated with G has the arcs (i, j) and (j, i) whenever $ij \in E$. Given a directed graph $D = (V, A)$ and $S \subseteq V$, we denote by $\delta^+(S)$ (respectively, $\delta^-(S)$) the set of arcs that enters (respectively, leaves) S. We write *cycle* instead of directed cycle.

This paper is organized as follows. Section 2 is devoted to the composition of polyhedra; §3 deals with the composition of facets; in §4 we study the algorithmic aspects of this composition; in §5 we study compositions of TDI systems; §6 is dedicated to the study of the orientations of K_5; in §7 we study graphs with no $K_{3,3}$ minor.

2. Compositions of polyhedra. In this section, we assume that $D = (V, A)$ is a connected digraph having a two-node cutset $\{u, v\}$, i.e.,

(i) $V = V_1 \cup V_2$,
(ii) $V_1 \cap V_2 = \{u, v\}$,
(iii) $D \backslash \{u, v\}$ is disconnected.

For $k = 1, 2$, we define $\bar{D}_k = (V_k, \bar{A}_k)$, where

$$\bar{A}_k = A(V_k) \cup \{(u, v), (v, u)\},$$

and $\bar{D} = (V, \bar{A})$ with $\bar{A} = A \cup \{(u, v), (v, u)\}$. Note that we could create parallel arcs in this way; this is not a problem in any of the treatments that follow.

We explain how to describe $P(D)$ from systems defining $P(\bar{D}_1)$ and $P(\bar{D}_2)$.

The inequality

$$x(u, v) + x(v, u) \leq 1$$

defines a facet of $P(\bar{D})$; it is easy to see that this is the only facet-defining inequality that has nonzero coefficients for both $x(u, v)$ and $x(v, u)$. Therefore the facets of $P(\bar{D}_k)$, for

$k = 1, 2$, can be classified into the five types below:

$$\text{(2.1a)} \qquad \sum_{(i,j)\in A_k} a_l^k(i,j)x(i,j) \le \alpha_l^k, \qquad l \in I_1^k,$$

$$\text{(2.1b)} \qquad \sum_{(i,j)\in A_k} a_l^k(i,j)x(i,j) + x(u,v) \le \alpha_l^k, \qquad l \in I_2^k,$$

$$\text{(2.1c)} \qquad \sum_{(i,j)\in A_k} a_l^k(i,j)x(i,j) + x(v,u) \le \alpha_l^k, \qquad l \in I_3^k,$$

$$\text{(2.1d)} \qquad x(u,v) + x(v,u) \le 1,$$

$$\text{(2.1e)} \qquad x(i,j) \ge 0 \quad \text{for } (i,j) \in \bar{A}_k,$$

where I_1^k is the set of inequalities with zero coefficients for $x(u, v)$ and $x(v, u)$; I_2^k is the set of inequalities with a nonzero coefficient for $x(u, v)$ and a zero coefficient for $x(v, u)$; I_3^k is the set of inequalities having a zero coefficient for $x(u, v)$ and a nonzero coefficient for $x(v, u)$.

The equation

$$x(u, v) + x(v, u) = 1$$

defines a facet $F(\bar{D}_k)$ of $P(\bar{D}_k)$ and a facet $F(\bar{D})$ of $P(\bar{D})$. The polytope $P(D)$ is a projection of $F(\bar{D})$ along the variables $x(u, v)$ and $x(v, u)$. The following theorem lets us find a system that describes $F(\bar{D})$.

THEOREM 2.1. *The polytope $F(\bar{D})$ is defined by the union of the systems that define $F(\bar{D}_1)$ and $F(\bar{D}_2)$.*

Proof. Let Q denote the polytope defined by the union of these two systems. Let x be a vector in Q. Let x_1 (respectively, x_2) be the restriction of x to $\bar{A}_1$ (respectively, $\bar{A}_2$); we have that

$$x_1 = \sum \alpha_i y_i, \quad \alpha_i \ge 0, \quad \sum \alpha_i = 1, \quad \text{and}$$

$$x_2 = \sum \beta_i z_i, \quad \beta_i \ge 0, \quad \sum \beta_i = 1,$$

where $\{y_i\}$ and $\{z_i\}$ are integer vectors in $F(\bar{D}_1)$ and $F(\bar{D}_2)$, respectively.

Since

$$\sum \{\alpha_i \,|\, y_i(u, v) = 1\} = x(u, v) = \sum \{\beta_i \,|\, z_i(u, v) = 1\},$$

$$\sum \{\alpha_i \,|\, y_i(v, u) = 1\} = x(v, u) = \sum \{\beta_i \,|\, z_i(v, u) = 1\},$$

we can match vectors in $\{y_i\}$ with vectors in $\{z_i\}$ to write x as a convex combination of integer vectors in $F(\bar{D})$. □

We now need a way to project $x(u, v)$ and $x(v, u)$; this is given by the following result of Balas and Pulleyblank [2].

THEOREM 2.2. *Let $Z = \{(w, x) \,|\, Aw + Bx \le b,\ w \ge 0,\ x \ge 0\}$; the projection of Z along the subspace of the w variables is*

$$X = \{x \,|\, (vB)x \le vb,\ \forall v \in \text{extr } Y,\ x \ge 0\},$$

where extr Y *denotes the set of extreme rays of*

$$Y = \{y \,|\, yA \ge 0\}.$$

In our case, the rows of the matrix A are of the types below:

$$\begin{array}{cc} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{array}.$$

By observing the extreme rays of the set Y, we can deduce the next theorem.

THEOREM 2.3. *The polytope $P(D)$ is defined by the inequalities* (2.1a), *together with* $x(i,j) \geq 0$ *and the mixed inequalities*

$$\sum_{(i,j)\in A_1} a_1^1(i,j)x(i,j) + \sum_{(i,j)\in A_2} a_p^2(i,j)x(i,j) \leq \alpha_l^1 + \alpha_p^2 - 1 \tag{2.2}$$

for $(l,p) \in (I_2^1 \times I_3^2) \cup (I_3^1 \times I_2^2)$.

COROLLARY 2.4. *If $P(\bar{D}_1)$ and $P(\bar{D}_2)$ are defined by* (1.1), (1.2), *then $P(D)$ is, also.*

A constraint with coefficients 0 or 1 is called a *rank* inequality. It also follows from Theorem 2.3 that, if $P(\bar{D}_1)$ and $P(\bar{D}_2)$ are defined by rank inequalities, then $P(D)$ is, also.

When we add the arcs (u, v) and (v, u), we could create parallel arcs in $\bar{D}_1$ and $\bar{D}_2$. If this is the case, for every inequality of type (2.1b) or (2.1c), we would have a similar inequality in (2.1a). This observation and Theorem 2.3 imply the next corollary.

COROLLARY 2.5. *The polytope $P(\bar{D})$ is defined by* (2.1) *and* (2.2).

3. Compositions of facets. The purpose of this section is to prove that Theorem 2.3 gives a minimal description of $P(D)$.

LEMMA 3.1. *Inequalities* (2.1b) *and* (2.1c) *define facets of $F(\bar{D}_k)$.*

Proof. Consider (2.1b). Let $ax \leq \alpha$ be one of them and let $S = \{x_1, \ldots, x_p\}$ be the set of extreme points of $P(\bar{D}_k)$ that satisfy $ax = \alpha$. There are $|\bar{A}_k|$ linearly independent vectors in S.

Since dim $(F(\bar{D}_k)) = |\bar{A}_k| - 1$, we need the same number of linearly independent vectors that satisfy $ax = \alpha$. Let x_r be a vector in S; if $x_r \in F(\bar{D}_k)$, we set $x'_r = x_r$; otherwise, we set $x'_r(i,j) = x(i,j)$ for $(i,j) \neq (v,u)$, and $x'_r(v,u) = 1$. Since we have modified only one component of the vectors in S, the set $S' = \{x'_1, \ldots, x'_p\}$ contains $|\bar{A}_k| - 1$ linearly independent vectors in $F(\bar{D}_k)$ that satisfy $ax = \alpha$. □

Because of the structure of system (2.1) and the lemma above, we can see that, for any inequality (2.1b) or (2.1c), there is no other constraint in (2.1) that defines the same face of $F(\bar{D}_k)$.

THEOREM 3.2. *Inequalities* (2.2) *define facets of $P(D)$.*

Proof. Assume $(l,p) \in I_2^1 \times I_3^2$. We show that there exists a vector $\bar{x} \in P(D)$ that satisfies this inequality as equation and all others as strict inequality.

For the inequality

$$\sum_{(i,j)\in \bar{A}_1} a_l^1(i,j)x(i,j) + x(u,v) \leq \alpha_l^1, \qquad l \in I_2^1,$$

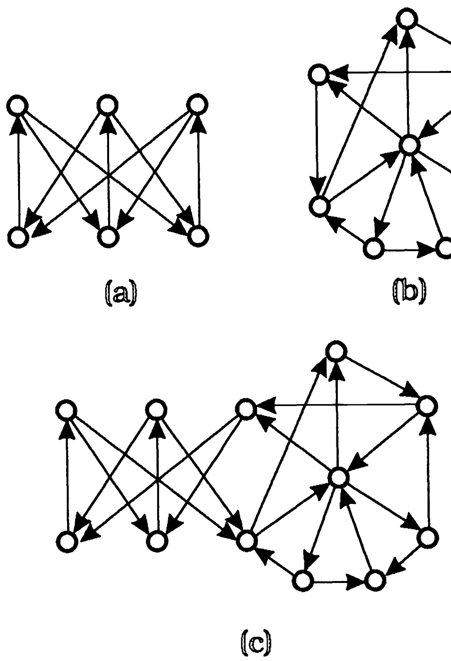
let $S = \{x_1, \ldots, x_r\}$ be the set of extreme points of $F(\bar{D}_1)$ that satisfy it as equation.

For

$$\sum_{(i,j)\in \bar{A}_2} a_p^2(i,j)x(i,j) + x(v,u) \leq \alpha_p^2, \qquad p \in I_3^2,$$

let $T = \{y_1, \ldots, y_s\}$ be the extreme points of $F(\bar{D}_2)$ that satisfy it as equation.

Each vector x_t can be matched with a vector $y_{t'}$ to give a vector in $F(\bar{D})$. Let $\{z_1, \ldots, z_d\}$ be the set thus obtained. Let z'_t be the vector obtained by deleting the



(a)

(b)

(c)

FIG. 3.1

components $z'_t(u, v)$ and $z'_t(v, u)$ from z_t. Then

$$\bar{x} = \frac{1}{d}(z'_1 + \cdots + z'_d)$$

is the required vector. □

We conclude this section with one example of this composition of facets. Denote by $D_1 = (V_1, A_1)$ and $D_2 = (V_2, A_2)$ the graphs in Fig. 3.1(a) and 3.1(b), respectively. Let $D = (V, A)$ be the graph in Fig. 3.1(c). Let $\bar{D}_1$ and $\bar{D}_2$ be defined as in §2. Grötschel, Jünger, and Reinelt [7] proved that

$$\sum_{(i,j)\in A_1} x(i,j) \le 7 \tag{3.1}$$

and

$$\sum_{(i,j)\in A_2} x(i,j) \le 11$$

define facets of $P(D_1)$ and $P(D_2)$, respectively; these inequalities also define facets of $P(\bar{D}_1)$ and $P(\bar{D}_2)$, respectively. Theorem 3.2 implies that

$$\sum_{(i,j)\in A} x(i,j) \le 17$$

defines a facet of $P(D)$.

4. Algorithmic aspects. The problem of optimizing a linear function over $P(D)$ can be decomposed in a similar way as shown in this section.

Let $w : \bar{A} \twoheadrightarrow \Re_+$ be a weight function. To simplify the notation, we denote by a_1 (respectively, a_2) the arc (u, v) (respectively, (v, u)). Let ω_i^j be the maximum weight of an acyclic subgraph of $\bar{D}_i$ that contains a_j.

Define

$$\alpha = \omega_1^1 - \omega_1^2, \quad \chi = \max\{0, \alpha\}, \quad \kappa = \chi - \alpha.$$

Let $\sigma \in \Re$ such that

$$\chi = \omega_1^1 - \sigma, \qquad \kappa = \omega_1^2 - \sigma. \tag{4.1}$$

Define $w' : \bar{A}_2 \twoheadrightarrow \Re_+$ as

$$w'(i,j) = w(i,j) \quad \text{for } (i,j) \in A_2,$$
$$w'(a_1) = \chi,$$
$$w'(a_2) = \kappa.$$

THEOREM 4.1. *The maximum weight of an acyclic subgraph of $\bar{D}$ with respect to w is $\lambda + \sigma$, where λ is the maximum weight of an acyclic subgraph of $\bar{D}_2$ with respect to w'.*

Proof. The maximum weight of an acyclic subgraph of D is

$$\max\{\omega_1^1 + \omega_2^1 - w(a_1), \omega_1^2 + \omega_2^2 - w(a_2)\};$$

it follows from (4.1) that this is equal to

$$\sigma + \max\{\chi + \omega_2^1 - w(a_1), \kappa + \omega_2^2 - w(a_2)\} = \sigma + \lambda. \qquad \square$$

5. Total dual integrality. A system $Ax \leq b$ is called *total dual integral* (TDI) if the dual problem of

$$\max wx \quad \text{s.t. } Ax \leq b$$

has an integer optimal solution for every integer vector w such that the maximum exists. If the system is TDI and b is an integer vector, then $\{x | Ax \leq b\}$ has integer extreme points. In this section, we prove that, if systems (2.1) are TDI, then the system given by Corollary 2.5 is also TDI. For $k = 1, 2$, we denote by $\mathscr{P}_k$ the linear program

$$\max wx \quad \text{s.t. (2.1).}$$

The dual problem of $\mathscr{P}_k$ is $\mathscr{D}_k$,

$$\text{minimize} \sum_{p=1,2,3} \sum_{l \in I_p^k} y_l^k \alpha_l^k + y_0^k$$

$$\text{s.t.} \quad \sum_{p=1,2,3} \sum_{l \in I_p^k} a_l^k(i,j) y_l^k \geq w(i,j) \quad \text{for } (i,j) \in A_k,$$

$$\sum_{l \in I_2^k} y_l^k + y_0^k \geq w(u, v),$$

$$\sum_{l \in I_3^k} y_l^k + y_0^k \geq w(v, u),$$

$$y_l^k \geq 0, \qquad y_0^k \geq 0.$$

We denote by $\mathscr{P}_3$ the linear program

$$\max wx \quad \text{s.t. (2.1) and (2.2).}$$

Below is the dual $\mathscr{D}_3$.

$$\text{minimize} \sum_{q=1,2}\ \sum_{p=1,2,3}\ \sum_{l\in I_p^q} y_l^q\alpha_l^q + \sum_{(k,l)\in(I_2^1\times I_3^2)\cup(I_3^1\times I_2^2)} (\alpha_k^1+\alpha_l^2-1)z_{k,l} + y_0$$

$$\text{s.t.} \quad \sum_{p=1,2,3}\ \sum_{l\in I_p^1} a_l^1(i,j)y_l^1 + \sum_{(k,l)\in(I_2^1\times I_3^2)\cup(I_3^1\times I_2^2)} a_k^1(i,j)z_{k,l} \geq w(i,j)$$

for $(i,j)\in A_1$,

$$\sum_{p=1,2,3}\ \sum_{l\in I_p^2} a_l^2(i,j)y_l^2 + \sum_{(k,l)\in(I_2^1\times I_3^2)\cup(I_3^1\times I_2^2)} a_l^2(i,j)z_{k,l} \geq w(i,j)$$

for $(i,j)\in A_2$,

$$\sum_{q=1,2}\ \sum_{l\in I_2^q} y_l^q + y_0 \geq w(u,v),$$

$$\sum_{q=1,2}\ \sum_{l\in I_3^q} y_l^q + y_0 \geq w(v,u),$$

$$y_l^q \geq 0, \quad y_0 \geq 0, \quad z_{k,l}\geq 0.$$

Suppose that systems (2.1) are TDI and that the weights w are integer. To prove that $\mathscr{D}_3$ has an integer optimal solution, we must study four cases; we present one of them; the others are similar.

Suppose that we apply the algorithm of §4. Let $\mathscr{A}_2$ be the maximum acyclic arc set in $\bar{D}_2$ with respect to w'. We consider the case where $\alpha\geq 0$ and $(v,u)\notin\mathscr{A}_2$. Thus $\chi=\alpha$, $\kappa=0$. We can assume that $(u,v)\in\mathscr{A}_2$. There is an integer optimal solution $\bar{y}^2$ of $\mathscr{D}_2$. Complementary slackness implies that

$$\sum_{l\in I_2^2} \bar{y}_l^2 + \bar{y}_0^2 = \alpha.$$

Now let us associate the weights w_1 to the arcs in $\bar{D}_1$, where

$$w_1(i,j) = \begin{cases} w(i,j) & \text{if } (i,j)\neq(v,u),\\ w(v,u)+\alpha & \text{if } (i,j)=(v,u).\end{cases}$$

Since w_1 is integer, there is an integer vector $\bar{y}^1$ that is an optimal solution of $\mathscr{D}_1$. Thus we have

$$\sum_{l\in I_3^1} \bar{y}_l^1 + \bar{y}_0^1 \geq w(v,u)+\alpha,$$

which implies that

$$\sum_{l\in I_3^1} \bar{y}_l^1 + \bar{y}_0^1 \geq \alpha.$$

Suppose that $\bar{y}_0^1<\alpha$ (the case where $\bar{y}_0^1\geq\alpha$ is similar). There is a set $\mathscr{I}\subseteq I_3^1$ such that

$$\sum_{l\in\mathscr{I}} \bar{y}_l^1 + \bar{y}_0^1 < \alpha$$

and

$$\sum_{l\in\mathscr{I}} \bar{y}_l^1 + \bar{y}_s^1 + \bar{y}_0^1 \geq \alpha,$$

where $s \in I_3^1 \setminus \mathcal{I}$. Let $\bar{\mathcal{I}} = \mathcal{I} \cup \{s\}$. Define $t = \alpha - (\sum_{l \in \mathcal{I}} \bar{y}_l^1 + \bar{y}_0^1)$. Consider now the system of equations

$$\sum_{k \in \bar{\mathcal{I}}} z_{k,l} + \gamma_l = \bar{y}_l^2 \quad \text{for } l \in I_2^2,$$

$$\sum_{k \in \bar{\mathcal{I}}} \rho_k + \delta = \bar{y}_0^2,$$

$$\sum_{l \in I_2^2} z_{k,l} + \rho_k = \begin{cases} \bar{y}_k^1 & \text{for } k \in \mathcal{I}, \\ t & \text{for } k = s, \end{cases}$$

$$\sum_{l \in I_2^2} \gamma_l + \delta = \bar{y}_0^1.$$

This matrix is totally unimodular; actually, it is a network flow matrix. Therefore the above system has a nonnegative integer solution.

Now consider the vector defined below:

$$\tilde{y}_i^1 = \begin{cases} \bar{y}_i^1 & \text{for } i \in I_1^1 \cup I_2^1 \cup (I_3^1 \setminus \bar{\mathcal{I}}), \\ \rho_i & \text{for } i \in \mathcal{I}, \\ \bar{y}_s^1 - t + \rho_s & \text{if } i = s; \end{cases}$$

$$\tilde{y}_i^2 = \begin{cases} \bar{y}_i^2 & \text{for } i \in I_1^2 \cup I_3^2, \\ \gamma_i & \text{for } i \in I_2^2; \end{cases}$$

$$\tilde{y}_0 = \delta,$$

$$\tilde{z}_{k,l} = \begin{cases} z_{k,l} & \text{for } k \in \bar{\mathcal{I}},\ l \in I_2^2, \\ 0 & \text{otherwise.} \end{cases}$$

The vector $(\tilde{y}, \tilde{z})$ is a feasible solution of $\mathcal{D}_3$ and its value is $\lambda + \sigma$ as defined in Theorem 4.1.

The remaining cases can be treated in a similar way, they are

(1) $\alpha \geq 0$ and $a_2 \in \mathcal{A}_2$,
(2) $\alpha < 0$ and $a_2 \notin \mathcal{A}_2$,
(3) $\alpha < 0$ and $a_2 \in \mathcal{A}_2$.

Therefore the system defined in Corollary 2.5 is TDI.

6. Orientations of K_5.

In this section, we prove that system (1.1), (1.2) when associated with $D(K_5)$ is TDI. This has been conjectured by Jünger [8].

Define the linear program

$$\begin{aligned} &\text{maximize } wx \\ (6.1) \qquad &\text{s.t.} \quad \sum_{(i,j) \in C} x(i,j) \leq |C| - 1 \quad \text{for every directed cycle } C, \\ &0 \leq x(i,j) \leq 1 \quad \text{for every arc } (i,j) \end{aligned}$$

and its dual

$$\text{(6.2)} \qquad \begin{aligned} &\text{minimize } \sum y_C(|C|-1) + \sum \gamma_{(i,j)} \\ \text{s.t.} \quad &\sum_{C \in \mathscr{C}_{(i,j)}} y_C + \gamma_{(i,j)} \geq w(i,j) \quad \text{for each arc } (i,j), \\ &y \geq 0,\ \gamma \geq 0. \end{aligned}$$

Here $\mathscr{C}_{(i,j)}$ denotes the set of cycles that contain (i,j).

Let us denote by $\mathscr{P}$ and $\mathscr{D}$ problems (6.1) and (6.2) when they are associated with $D(K_5)$. We construct $D' = (V', A')$ as follows:

(a) If $w(i,j) > w(j,i) \geq 0$, then $(i,j) \in A'$ and $w'(i,j) = w(i,j) - w(j,i)$;

(b) If $w(i,j) = w(j,i) > 0$, then $(i,j) \in A'$ or $(j,i) \in A'$ but not both, say $(i,j) \in A'$ with $w'(i,j) = 0$;

(c) If $w(i,j) \geq 0 > w(j,i)$, then $(i,j) \in A'$ and $w'(i,j) = w(i,j)$;

(d) If $w(i,j) < 0$ and $w(j,i) < 0$, we do not put any arc between i and j.

Denote by $\mathscr{P}'$ and $\mathscr{D}'$ problems (6.1) and (6.2) when they are associated with D'. Let $\bar{x}$ and $(\bar{y}, \bar{\gamma})$ be optimal solutions of $\mathscr{P}'$ and $\mathscr{D}'$, respectively; we can construct optimal solutions of $\mathscr{P}$ and $\mathscr{D}$ as shown below.

Set $\bar{\bar{x}}(i,j) = \bar{x}(i,j)$, $\bar{\bar{x}}(j,i) = 1 - \bar{x}(i,j)$ if (a) or (b) holds; $\bar{\bar{x}}(i,j) = \bar{x}(i,j)$, $\bar{\bar{x}}(j,i) = 0$ if (d) holds; $\bar{\bar{x}}(i,j) = \bar{\bar{x}}(j,i) = 0$ if (d) holds; $\bar{\bar{y}}_C = \bar{y}_C$ if C is a cycle of D'; $\bar{\bar{y}}_C = \bar{\gamma}(i,j) + w(j,i)$ if $C = \{(i,j),(j,i)\}$ and (a) or (b) holds; $\bar{\bar{y}}_C = \bar{\gamma}(i,j)$ if $C = \{(i,j), (j,i)\}$ and (c) holds; $\bar{\bar{y}}_C = 0$ otherwise; and $\bar{\bar{\gamma}}_{(i,j)} = 0$ for all (i,j). Therefore, instead of studying $D(K_5)$, we study the different orientations of K_5.

In the remainder of this section, $D = (V, A)$ denotes an orientation of K_5. We first prove that (6.1) defines a polytope with integral extreme points. Let $\bar{x}$ be an extreme point; the following remarks allow us to rule out many cases.

Remark 6.1. If there is an arc (i,j) that does not belong to any cycle, then $\bar{x}$ is integer-valued.

Proof. If (i,j) does not belong to any cycle, then the nontrivial inequalities of (6.1) are associated with $D\backslash(i,j)$, and this is a planar graph. □

Remark 6.2. If $\bar{x}(i,j) = 0$ for an arc (i,j), then $\bar{x}$ is integer-valued.

Proof. Consider $D' = D\backslash(i,j)$ and let x' be $\bar{x}$ without the component associated with (i,j). Since D' is planar, we have that $x' \in P(D')$ and x' is a convex combination of a set of vectors $\{x_j\}$ incidence vectors of acyclic subgraphs of D'. For each vector x_j, we can add a zero component and obtain the incidence vector of an acyclic subgraph of D. We have then that $\bar{x}$ is a convex combination of them. □

Remark 6.3. For any set S, $\emptyset \neq S \subset V$, there is at least one arc $(i,j) \in \delta^+(S) \cup \delta^-(S)$ with $\bar{x}(i,j) = 1$.

Proof. If $0 < \bar{x}(i,j) < 1$ for every arc $(i,j) \in \delta^+(S) \cup \delta^-(S)$, define x' as

$$x'(i,j) = \begin{cases} \bar{x}(i,j) + \varepsilon & \text{if } (i,j) \in \delta^+(S), \\ \bar{x}(i,j) - \varepsilon & \text{if } (i,j) \in \delta^-(S), \\ \bar{x}(i,j) & \text{otherwise.} \end{cases}$$

For ε sufficiently small, x' satisfies (6.1), and, if some of these inequalities hold as equation for $\bar{x}$, they also do for x'. This contradicts the assumption that $\bar{x}$ is an extreme point. □

Remark 6.4. It follows from Remark 6.1 that we can assume that D is strongly connected.

Remark 6.5. It follows from Remark 6.3 that we can assume that there is a tree $\mathcal{T}$ of arcs (i, j) with $\bar{x}(i, j) = 1$.

Remark 6.6. For every variable $x(i, j)$ with $0 < \bar{x}(i, j) < 1$, we can assume that it appears in at least two tight cycle constraints.

Proof. If $0 < \bar{x}(i, j) < 1$ and $x(i, j)$ appears in only one tight constraint, we can set $x(i, j) = 0$; this new vector is also an extreme point, and, from Remark 6.2, we can conclude that it is integral. This gives a contradiction. □

Remark 6.7. If there is a node v that covers every cycle, then $\bar{x}$ should be integer-valued.

Proof. Consider $D' = (V', A')$, where $V' = V\backslash\{v\} \cup \{s, t\}$ and A' is defined below:

$$(i, j) \in A, \quad i \neq v, \quad j \neq v \Rightarrow (i, j) \in A',$$

$$(v, i) \in A \Rightarrow (s, i) \in A',$$

$$(i, v) \in A \Rightarrow (i, t) \in A'.$$

Let M be the incidence matrix of all directed paths from s to t in D'; it is well known that, for any $w \geq 0$, the problem

$$\text{minimize } wx \quad \text{s.t. } Mx \geq 1, x \geq 0$$

has an integer-valued optimal solution $\tilde{x}$. Since M is also the incidence matrix of the cycles in D, the vector $\hat{x}$ defined by

$$\hat{x}(i, j) = 1 - \tilde{x}(i, j) \quad \text{for } (i, j) \in A$$

is an optimal solution of (6.1). □

Now we can prove the following result.

THEOREM 6.8. *All the extreme points of the polyhedron defined by* (6.1) *are integral.*

Proof. There are three cases to study.

Case 1. The tree $\mathcal{T}$ contains a directed path with two arcs; i.e., suppose that $\bar{x}(1, 2) = \bar{x}(2, 3) = 1$. We have that $(1, 3) \in A$. Consider the cycle inequalities that are tight for $\bar{x}$; if there is a cycle that contains (1, 3) and goes through 2, it should be $C = (1, 3, 4, 2, 5, 1)$. Since

$$\bar{x}(2, 3) + \bar{x}(3, 4) + \bar{x}(4, 2) \leq 2,$$

$$\bar{x}(2, 5) + \bar{x}(5, 1) + \bar{x}(1, 2) \leq 2,$$

$$\bar{x}(1, 3) \leq 1,$$

we have that

$$\bar{x}(1, 3) + \bar{x}(3, 4) + \bar{x}(4, 2) + \bar{x}(2, 5) + \bar{x}(5, 1) \leq 3,$$

a contradiction.

Therefore we should assume that every cycle containing (1, 3) does not go through 2.

Let C be a cycle that contains (1, 3); if the constraint associated with C is tight, then $\bar{x}(1, 3) = 1$; otherwise, the constraint

$$\sum_{(i,j) \in C'} x(i, j) \leq |C'| - 1$$

would be violated, where $C' = C\backslash\{(1, 3)\} \cup \{(1, 2), (2, 3)\}$. Then we must only study the case where $\bar{x}(1, 3) = 1$.

Now consider the graph $D' = D\backslash(1, 3)$. Let $\bar{x}'$ be the restriction of $\bar{x}$ to the arc set of D'. We have that $\bar{x}' = \sum \lambda_i y_i$, $\lambda_i \geq 0$, $\sum \lambda_i = 1$, where the vectors $\{y_i\}$ are incidence vectors of acyclic subgraphs of D'.

Now define $\bar{y}_i$ as follows:

$$\bar{y}_i(k, l) = y_i(k, l) \quad \text{if } (k, l) \neq (1, 3),$$
$$\bar{y}_i(1, 3) = 1 \quad \text{for all } i.$$

The vectors $\{\bar{y}_i\}$ are incidence vectors of acyclic subgraphs of D, and $\bar{x} = \sum \lambda_i \bar{y}_i$; then $\bar{x}$ should be an integer vector.

Case 2. The tree $\mathscr{T}$ is $\{(1, 2), (3, 2), (3, 4), (3, 5)\}$. We can assume that $(4, 5) \in A$. We should assume that $(1, 3) \in A$; otherwise, there is no cycle going through 3. We also assume that $(5, 1) \in A$; otherwise, 4 covers every cycle. For the same reasons, we assume that $(4, 1) \in A$. Also, we should have that $(5, 2)$ and $(2, 4)$ are in A; otherwise, 1 would cover every cycle. See Fig. 6.1.

We have that

$$\begin{aligned} &\bar{x}(2, 4) + \bar{x}(4, 1) + \bar{x}(1, 3) \leq 2, \\ &\bar{x}(2, 4) + \bar{x}(4, 1) + \bar{x}(1, 3) + \bar{x}(5, 2) \leq 3. \end{aligned} \tag{6.3}$$

Thus (6.3) holds as equation only if $\bar{x}(5, 2) = 1$; then, however, we would be in Case 1. If (6.3) does not hold, we would have that (5, 2) only appears in one tight cycle inequality; then we can apply Remark 6.6.

Case 3. The tree $\mathscr{T}$ is $\{(1, 2), (3, 2), (3, 4), (5, 4)\}$. We can assume that $(2, 4) \in A$. Therefore $(4, 1) \in A$; otherwise, there is no cycle going through 4. Then every cycle containing (2, 4) also contains (4, 1). Consider the cycle $C = \{(2, 4), (4, 1), (1, 2)\}$; we have that

$$\bar{x}(2, 4) + \bar{x}(4, 1) \leq 1.$$

Thus, if there is any other tight cycle inequality containing (2, 4), all its variables different from $\bar{x}(2, 4)$ and $\bar{x}(4, 1)$ should take the value 1; this is Case 1. This concludes Case 3 and the proof of the theorem. □

It remains to prove that (6.1) defines a TDI system. This has been proved by Applegate, Cook, and McCormick [1] using an algorithm that tests whether a system is TDI. We present here a proof that does not involve computer calculations.

Suppose that, for every integer vector $w \leq \mu$, $w \neq \mu$, problem (6.1) has an integer dual solution and let $z(w)$ be its value. Now we study the weights $w = \mu$; we should

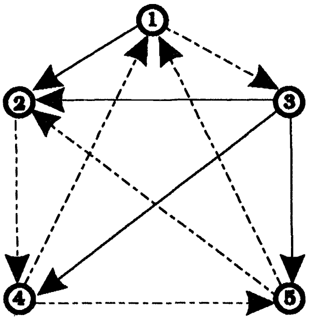


FIG. 6.1

assume that $\mu > 0$. If one component of μ is zero, then the associated arc can be removed yielding to a planar graph. Consider the set of inequalities that are tight for every optimal solution of (6.1).

Case 1. Assume that

$$x(i, j) \leq 1 \tag{6.4}$$

is tight; then we can define w' as

$$w'(k, l) = \begin{cases} w(k, l) & \text{if } (k, l) \neq (i, j), \\ w(i, j) - 1 & \text{if } (k, l) = (i, j). \end{cases}$$

We have that $z(w') = z(w) - 1$, and there is an integer dual solution for the objective function w'. We increase by 1 the value of the dual variable associated with (6.4) and we have a dual integer solution for the vector w.

Case 2. Consider now a cycle C of length 3, $C = \{(1, 2), (2, 3), (3, 1)\}$ say. Assume that the constraint associated with C is tight. Define w' by subtracting 1 from the costs coefficients of the arcs in C.

LEMMA 6.9. *The value of the new optimum is* $z(w') = z(w) - 2$.

Proof. If $z(w') = z(w)$, there is an optimal solution for the new objective function that does not contain any arc of C. However, we can always add one of the arcs of C to that solution without creating a cycle. This gives a solution for the original problem with value $z(w) + 1$, which is a contradiction.

If $z(w') = z(w) - 1$, there is an optimal solution for the new problem that contains one of the arcs of C. This is also an optimal solution for the objective function w, which is impossible because the constraint associated with C is tight for every optimum of the original problem. □

Given an integer optimal dual vector for the objective function w', we increase by 1 the value of the dual variable associated with C and we obtain a dual optimal solution for w.

Case 3. A cycle of length 4 is tight. In this case, it is easy to see that there is always a cycle of length 3 that is tight, and we are in Case 2.

Case 4. A cycle of length 5 is tight. In this case, there is also a cycle of length 3 or 4 that is tight.

We can then state the main result of this section.

THEOREM 6.10. *For* $D(K_5)$, *system* (6.1) *is* TDI.

7. Graphs with no $K_{3,3}$ minor. Grötschel, Jünger, and Reinelt [7] proved that, if D is a subdivision of the graph of Fig. 3.1(a), then an inequality analogous to (3.3) defines a facet of $P(D)$. Therefore, if G contains $K_{3,3}$ as a minor, then system (1.1), (1.2) is not sufficient to define $P(D(G))$. Wagner [10] proved that, if an undirected graph G has no $K_{3,3}$ minor, then either it is planar, it is K_5, or it has a two-vertex cut. We know that, if G is planar or it is K_5 then $P(D(G))$ is defined by (1.1), (1.2), and this is a TDI system. This and the result of §5 imply the following result.

THEOREM 7.1. *Given a graph G, system* (1.1), (1.2) *defines* $P(D(G))$ *and is* TDI *if and only if G does not contain* $K_{3,3}$ as a minor.

An immediate consequence is the result below.

COROLLARY 7.2. *If D is a directed graph that does not contain a subdivision of* $K_{3,3}$ *then the cardinality of a minimum feedback set is equal to the maximum number of arc-disjoint cycles.*

For planar graphs, this follows from the theorem of Lucchesi and Younger [9].

Let A be a matrix with nonnegative entries. Let $P = \{x | Ax \geq 1, x \geq 0\}$ and let us assume that this system is not redundant. If B is the matrix whose rows are the extreme points of P, then the pair (A, B) is called a blocking pair; see Fulkerson [4]. Actually, extreme points of $Q = \{x | Bx \geq 1, x \geq 0\}$ are the rows of A.

Theorem 7.1 implies the next corollary.

COROLLARY 7.3. *Given an undirected graph G, let A be the incidence matrix of the directed cycles of $D(G)$ and let B be the incidence matrix of all minimal feedback sets of $D(G)$. We have that (A, B) is a blocking pair if and only if G has no $K_{3,3}$ minor.*

REFERENCES

[1] D. L. APPLEGATE, W. COOK, AND S. T. MCCORMICK, *Integral infeasibility and testing total dual integrality*, Oper. Res. Lett., 10 (1991), pp. 37–41.

[2] E. BALAS AND W. R. PULLEYBLANK, *The perfectly matchable subgraph polytope of a bipartite graph*, Networks, 13 (1983), pp. 495–516.

[3] F. BARAHONA AND A. R. MAHJOUB, *Compositions in the Acyclic Subgraph Polytope*, Report No. 85371-OR, Institut für Ökonometrie und Operations Research, Universität Bonn, 1985.

[4] D. R. FULKERSON, *Blocking and anti-blocking polyhedra*, Math. Programming, 1 (1971), pp. 168–194.

[5] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[6] J. GRÖTSCHEL, M. JÜNGER, AND G. REINELT, *A cutting plane algorithm for the linear ordering problem*, Oper. Res., 32 (1984), pp. 1195–1220.

[7] ———, *On the acyclic subgraph polytope*, Math. Programming, 33 (1985), pp. 1–27.

[8] M. JÜNGER, *Polyhedral combinatorics and the acyclic subdigraph problem*, Ph.D. dissertation, University of Augsburg, 1983.

[9] C. L. LUCCHESI AND D. H. YOUNGER, *A minimax relation for directed graphs*, J. London Math. Soc., 17 (1978), pp. 369–374.

[10] K. WAGNER, *Uber eine erweiterung eines satzes von Kuratowski*, D. Math., 1937, pp. 280–285.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 403–418, August 1994

THE ALL-PAIRS MIN CUT PROBLEM AND THE MINIMUM CYCLE BASIS PROBLEM ON PLANAR GRAPHS*

DAVID HARTVIGSEN[†] AND RUSSELL MARDON[‡]

Abstract. The all-pairs min cut (APMC) problem on a nonnegative weighted graph is the problem of finding, for every pair of nodes, the min cut separating the pair. It is shown that on planar graphs, the APMC problem is equivalent to another problem, the minimum cycle basis (MCB) problem, on the dual graph. This is shown by characterizing the structure of MCBs on planar graphs in several ways. This leads to a new algorithm for solving both of these problems on planar graphs. The complexity of this algorithm equals that of the best algorithm for either problem, but is simpler.

Key words. networks flows, min cuts, cycle bases, graph theory

AMS subject classifications. 90, 05, 68

1. Introduction.

In this paper we consider two optimization problems on graphs: the all-pairs min cut (APMC) problem and the minimum cycle basis (MCB) problem. Our results are both structural and algorithmic. On the structural side, we show that these problems are "dual equivalent" on planar graphs (hence one problem can be transformed into the other in linear time) and we present several equivalent characterizations of their solutions. On the algorithmic side, we exploit these characterizations to present a new algorithm that solves both problems on planar graphs with nonnegative edge weights. Polynomial-time algorithms already exist for both problems, the fastest of which is for the APMC problem. The time complexity of our algorithm ($O(n^2 \log n + m)$, where n is the number of nodes and m is the number of edges) is equal to that of the fastest algorithm for the APMC problem, but our algorithm is simpler (it requires no complex data structures), and, in contrast, it explicitly solves the MCB problem.

In this section we present the two optimization problems. We follow this with a brief history of related work and then outline our results.

We consider graphs that are undirected, finite, and may contain loops and multiple edges. Let $G = (V, E)$ be a connected graph with edge weights and let $s, t \in V$. $C \subseteq E$ is called an *s–t cut* if $G' = (V, E \setminus C)$ contains no path from s to t and if C is minimal with respect to this property. The *weight of the s–t cut C* is the sum of the weights of its edges. The *all pairs min cut* (APMC) *problem* on G is the problem of finding a minimum s–t cut for every pair of nodes s and t in G. An *all-pairs minimum cut collection of* G, abbreviated APMC(G), is a minimal collection of cuts that solves the APMC problem on G.

Let us consider the second problem. A typical definition of a cycle in a graph is a minimal subgraph such that every node has degree 2. (These cycles are often referred to as "node simple.") However, we use the term "cycle" to refer to the edge sets of such subgraphs.

* Received by the editors January 24, 1990; accepted for publication (in revised form) April 28, 1993.

[†] Department of Management, College of Business Administration, University of Notre Dame, Notre Dame, Indiana 46556-5641.

[‡] Florida Agency for Health Care Administration, 325 John Knox Road, Tallahassee, Florida 32303.

To each cycle C in a graph $G = (V, E)$ we associate an incidence vector x, indexed on E, where $x_e = 1$ if $e \in C$ and $x_e = 0$, otherwise. A collection of cycles is called *independent* if their incidence vectors are independent over $GF(2)$. The vector space over $GF(2)$ generated by these vectors is called the *cycle space* of G. A maximal independent collection of cycles is called a *cycle basis.*

Consider a graph G with edge weights. The *weight of a cycle* is the sum of the weights of its edges, and the *weight of a collection of cycles* is the sum of the weights of its cycles. A *minimum cycle basis of* G, abbreviated MCB(G), is a cycle basis of minimum weight. The *minimum cycle basis* (MCB) *problem* is the problem of finding an MCB(G).

In this paper, we consider these two problems on planar graphs. We next survey the history of these problems, which is almost entirely algorithmic. Let us begin with the APMC problem.

Gomory and Hu (see [11], [9], and [15]) first considered the APMC problem. (For an application to communication networks, see [16].) They showed that these cuts can be found by solving just $n - 1$ min cut (or max flow) problems. Note that the APMC problem is NP-hard if negative weights are allowed, since an APMC of a graph contains a min cut in the graph. However, when edge weights are nonnegative, this method takes polynomial time and is the fastest known method for solving this problem.

Quite a bit of work has been devoted to the related problems on planar graphs of finding max flows (see [12], [13], [17], and [19]), min cuts (see [27] and [10]), and APMCs when the edge weights are all 1 (see [28]). In particular, Reif [27] gave an $O(n(\log n)^2)$ algorithm for finding a single pair min cut in a planar graph, which uses Dijkstra's algorithm [7] as a subroutine. Frederickson [10] showed that the complexity of Reif's algorithm can be improved to $O(n \log n)$ when Frederickson's shortest path algorithm is used as a subroutine. Hence the fastest known algorithm for solving the APMC problem in planar graphs is a combination of the algorithms of Gomory and Hu, Reif, and Frederickson, which runs in $O(n^2 \log n)$ time. We observe that the last two algorithms require the use of fairly sophisticated data structures. (An algorithm in [28] for finding APMCs when the weights are all 1 takes $O(n^2(\log n)^2)$ time.)

Let us next consider the history of cycle bases and the MCB problem. Perhaps the best-known class of cycle bases comes from the work of Kirchhoff [21] in 1847 on electrical circuit theory. For a graph $G = (V, E)$, let $T = (V, E')$ be a maximal spanning forest of G. Then the unique cycles in $e \cup E'$ for each $e \in E \setminus E'$ constitute a (strictly fundamental) cycle basis for G.

Another well-known class of cycle bases appears in MacLane's (1937) characterization of planar graphs (see [24]). If G is a plane graph, then the cycles corresponding to the (interior) faces are a cycle basis for G.

A brief survey of applications of cycle bases appears in [14]. For example, applications occur in the areas of electrical circuit theory [21] and the analysis of algorithms [22] (see also [4], [5], [6], [8], [25], [26], and [29]).

Applications and the history of the MCB problem are also surveyed in [14]. (For an application in electrical circuit theory, see [3].) Another area of application is in structural engineering (see [1] and [20]), where minimum-weight cycle bases are computed to efficiently analyze the flexibility of structures. Important examples arise on both planar and nonplanar graphs.

The MCB problem is NP-hard if negative weights are allowed (an MCB must include the shortest cycle in the graph). However, when the weights are nonnegative, Horton showed in 1987 [14] that the MCB problem can be solved in polynomial time.

The idea of Horton's algorithm is to first produce a polynomial length list of cycles that must contain an MCB and to then find the MCB by applying the greedy algorithm (see, e.g., [23]) to this list. This is the only known polynomial algorithm for the MCB problem, and its time complexity is $O(m^3n)$, or $O(n^4)$ on simple planar graphs. The expensive part of this algorithm is the greedy step that requires repeated applications of Gaussian elimination.

The paper is organized as follows. In §2 we state and prove the main theorem, which is two equivalent characterizations of MCBs in planar graphs. It follows immediately from this theorem that solving the APMC problem on a planar graph is essentially equivalent to solving the MCB problem on its dual. This theorem also leads to an algorithm for solving the MCB problem in a simple planar graph (hence it solves the APMC problem in the dual). We present this algorithm in §4. Our algorithm is a modification of Horton's algorithm: The idea is to shorten the list of candidate cycles and then extract the MCB from this list without using Gaussian elimination. The complexity of our algorithm is $O(n^2 \log n)$, which is the same as the complexity of the fastest planar APMC algorithm described above. However, our algorithm is simpler—it does not require Frederickson's shortest path algorithm as a subroutine nor does it use any complex data structures. Our algorithm is faster than the straightforward application of Horton's algorithm, which is $O(n^4)$. In the final section, we show how our algorithm can be modified to work on general planar graphs in $O(n^2 \log n + m)$ time.

2. Preliminaries and the main theorem. In this section we state our main theorem, which consists of two equivalent characterizations of MCBs in planar graphs. We precede the statement of the theorem with some important definitions and follow it with a brief discussion of the relationship between the MCB and APMC problems. This allows us to discuss only the MCB problem in the remainder of the paper. Also, after reading the material presented in this section, the reader may skip to the algorithm in §4 (thus skipping the proof in §3).

Let us refer to a planar graph that has been embedded in the plane as a *plane graph*. A plane graph G divides the plane into maximal open connected sets of points that we refer to as the *regions of* G. Each cycle C of G divides the plane into two maximal open connected sets of points. Let interior(C) denote the bounded set. If R_1 and R_2 are regions of a plane graph G, then a *cycle* C *separates* R_1 *and* R_2 if $R_1 \subseteq \text{interior}(C)$ or $R_2 \subseteq \text{interior}(C)$, but not both. A cycle C of a plane graph G is called an *internal face* if C bounds a bounded region of G. For simplicity, we refer to "internal faces" as "faces."

A collection of cycles $\mathbf{C}$ in a plane graph is called *nested* if, for each pair $C, C' \in \mathbf{C}$, either

$$\begin{aligned} &\text{interior}(C) \subset \text{interior}(C'); \\ \text{or} \quad &\text{interior}(C') \subset \text{interior}(C); \\ \text{or} \quad &\text{interior}(C) \cap \text{interior}(C') = \emptyset. \end{aligned} \tag{2.1}$$

For a graph with arbitrary edge weights, we use the term *path* to refer to simple paths (no repeated nodes). We use $\subset$ and $\supset$ to denote proper inclusion and proper containment, respectively.

Let $\mathbf{C}$ be a collection of cycles and let $\mathbf{F_C}$ be the set of faces of G not in $\mathbf{C}$. We construct a directed graph $D_\mathbf{C} = D(\mathbf{C} \cup \mathbf{F_C}, A)$ as follows (note that xy denotes an arc directed from node x to node y): $xy \in \mathbf{A}$ if and only if interior$(x) \supset$ interior(y)

and there exists no $z \in \mathbf{C} \cup \mathbf{F_C}$ such that interior$(x) \supset$ interior$(z) \supset$ interior(y). Observe that $D_{\mathbf{C}}$ is acyclic but may or may not be a forest.

We use the following terminology in reference to $D_{\mathbf{C}}$. If $xy \in A$, then we call x a *parent* of y, and y a *child* of x. If nodes x and y have the same parent, then we call them *siblings.* A node of in-degree 1 is called a *leaf.*

The main theorem follows. We emphasize that, in this paper, cycles are "node simple." If we relax this assumption, then we must restrict ourselves to nonnegative weights in the statement of Theorem 2.1.

THEOREM 2.1. *Let G be a plane graph with edge weights and let $\mathbf{C}$ be a collection of nested cycles in G. Then the following are equivalent:*

(i) $\mathbf{C}$ *is an* MCB(G);

(ii) $\mathbf{C}$ *is a minimal set with the following property: For every pair of regions,* $\mathbf{C}$ *contains a minimum-weight cycle that separates the pair;*

(iii) $\mathbf{C}$ *is a minimum-weight collection of cycles that satisfies the following three conditions:*

(a) $D_{\mathbf{C}}$ *is a forest;*

(b) *Every nonleaf in* $D_{\mathbf{C}}$ *has a unique child in* $\mathbf{F_C}$;

(c) $D_{\mathbf{C}}$ *has no isolated node in* $\mathbf{F_C}$.

Let us now discuss the relationship of the MCB and APMC problems. If $G = (V, E)$ is a plane graph, let $G^d = (R, E)$ denote the dual plane graph of G, where R is the set of regions of G. (That is, G^d has a node associated with each region of G, and two nodes of G^d are connected by an edge if and only if the corresponding regions of G have a common edge in their boundaries.) Since there is a 1–1 correspondence between the edge sets of G and G^d, we denote both edge sets by E; hence $E' \subseteq E$ denotes a set of edges in both G and G^d. An edge weighting on G induces an edge weighting on G^d. For $C \subseteq E$, it is well known that

$$C \text{ is a cut in } G \text{ iff } C \text{ is a cycle in } G^d. \tag{2.2}$$

It is easy to see that if G is a connected plane graph, then there is a natural 1–1 correspondence between the nodes of G and the regions of G^d. (In particular, $G = (G^d)^d$.) It easily follows from (2.2) that if G is connected, C is an s–t cut in G if and only if $\mathbf{C}$ separates the regions corresponding to s and t in G^d. The following corollary follows immediately from these observations and the above theorem (i.e., (i) if and only if (ii)).

COROLLARY 2.2. *Let G be a connected plane graph with edge weights and let $\mathbf{C}$ be a collection of nested cycles in G^d. Then $\mathbf{C}$ is an* APMC(G) *if and only if $\mathbf{C}$ is an* MCB(G^d).

In §4 we easily show that every plane graph has a nested MCB. Thus it follows from the above corollary that we can find an APMC(G) by finding a nested MCB(G^d). (Since the dual of a simple plane graph can be found in $O(n)$ time, the complexity of this algorithm for finding an APMC(G) is the same as the complexity of finding an MCB(G). See Proposition 4.22 and Observation 4.20.) Section 4 contains an algorithm that finds a nested MCB of a simple planar graph. In the remainder of the paper, we address only the MCB problem.

3. Proof of Theorem 2.1. This section is devoted to proving Theorem 2.1. We do so in parts, introducing useful definitions and propositions as necessary.

PROPOSITION 3.1. *Let $\mathbf{C}$ be a collection of independent cycles in a plane graph G. Then $\mathbf{C}$ is a cycle basis for G if and only if the number of nonfaces in $\mathbf{C}$ equals the number of faces not in $\mathbf{C}$.*

Proof. The faces of a planar graph form a cycle basis, hence (the number of faces not in $\mathbf{C}$) + (the number of faces in $\mathbf{C}$) = $\dim(G)$. Equivalently, (the number of faces not in $\mathbf{C}$) + $[|\mathbf{C}|$ − (the number of nonfaces in $\mathbf{C}$)] = $\dim(G)$. The result follows. □

The following proposition is key to proving Theorem 2.1.

PROPOSITION 3.2. *Let* $\mathbf{C}$ *be a collection of cycles of a plane graph* G. *Then* $\mathbf{C}$ *is a nested cycle basis for* G *if and only if* $\mathbf{C}$ *satisfies items* (a)–(c) *in Theorem* 2.1.

Proof. ($\Rightarrow$) Since $\mathbf{C}$ is nested, $D_{\mathbf{C}}$ is a forest. Let x be a nonleaf node in $D_{\mathbf{C}}$. If x has no child in $\mathbf{F}_{\mathbf{C}}$, then x equals the sum of its children, contradicting our assumption that $\mathbf{C}$ is a cycle basis. So every nonleaf node has at least one child in $\mathbf{F}_{\mathbf{C}}$. If any nonleaf node has more than one child in $\mathbf{F}_{\mathbf{C}}$ or if $D_{\mathbf{C}}$ has an isolated node in $\mathbf{F}_{\mathbf{C}}$, then we contradict Proposition 3.1. The result follows.

($\Leftarrow$) Since $\mathbf{C}$ satisfies (a), it is nested; hence we must show that $\mathbf{C}$ is a basis. Since the number of cycles in $\mathbf{C}$ equals the number of faces of $G, \mathbf{C}$ has the correct cardinality. To show independence, it suffices to show that every face of G (the faces of G are a cycle basis) can be expressed as a sum of cycles in $\mathbf{C}$. Let C be a face of G not in $\mathbf{C}$. Then C is the sum of its parent and siblings in D_C, all of which are in $\mathbf{C}$. □

Proof that (i) *if and only if* (iii). This proof follows immediately from Proposition 3.2. □

For the remainder of the proof, we use the following simple propositions. For $G = (V, E)$, if $V' \subseteq V$, then the edges of E with exactly one node in V' are called the *coboundary of* V'.

PROPOSITION 3.3. *Let* $G = (V, E)$ *be a graph with edge weights. Let* $E' \subseteq E$ *be the coboundary of a node set* $V' \subseteq V$ *and let* E_1, E_2 *be a nontrivial partition of* E'. *If there exists a cycle* C *such that* $|E_1 \cap C|$ *and* $|E_2 \cap C|$ *are odd, then such a minimum-weight cycle is contained in every* MCB(G).

Proof. Let us color, with red, blue, and white, the edges in E_1, E_2, and $E \setminus E'$, respectively. Observe that since E' is the coboundary of a node set, each cycle intersects E' in an even number of edges; hence each cycle has the property that the numbers of red and blue edges in the cycle have the same parity. Let us call a cycle *odd* or *even* according to this parity. Observe next that

(3.1) no odd cycle can be expressed as a mod-2 sum of even cycles.

We can obtain an MCB(G) by applying the greedy algorithm to the cycles in G. (For a description of the greedy algorithm, see, e.g., [23].) By assumption, there exists an odd cycle, and by (3.1) we must, at some point of the greedy algorithm, first choose an odd cycle, call it C. Again by (3.1), no odd cycle can be expressed as a mod-2 sum of the cycles chosen before C. Hence C must be a minimum-weight odd cycle. The result follows. □

Let C_1 and C_2 be two cycles that bound regions of a plane graph G with edge weights (that is, C_1 and C_2 are faces or the cycle bounding the exterior region). A C_1–C_2 *path* is a path P with one endnode on $G(C_1)$, one endnode on $G(C_2)$, and no other node on $G(C_1)$ or $G(C_2)$. Let E' be the coboundary of the node set of such a path. Then there exists a cyclic ordering of the edges in E', say, $a, b, x_1, \dots, x_r, c, d, x_{r+1}, \dots, x_{r+t}$, where $a, b \in C_1$ and $c, d \in C_2$ (see Fig. 3.1). Note that it is possible that $b = c$ or $a = d$, however, $a \neq b$ and $c \neq d$. We call $\{b, x_1, \dots, x_r, c\}, \{d, x_{r+1}, \dots, x_{r+t}, a\}$ the (C_1, C_2)*-partition of* P. We immediately have the following proposition.

PROPOSITION 3.4. *Let* G *be a plane graph with edge weights. Let* C_1 *and* C_2 *be two cycles of* G *that bound regions* R_1 *and* R_2, *respectively, and let* P *be a* C_1–C_2

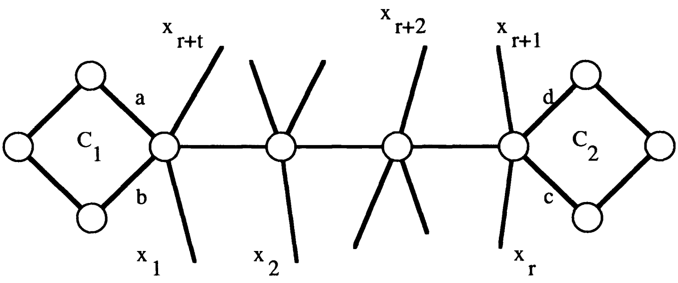


FIG. 3.1

path. *Then a cycle C separates R_1 and R_2 if and only if C intersects each set in the (C_1, C_2)-partition of P an odd number of times.*

PROPOSITION 3.5. *Let G be a plane graph with edge weights. If $\mathbf{C}$ is an* MCB(G), *then for every pair of regions in G there exists a cycle in $\mathbf{C}$ that separates the two regions and is a minimum weight such cycle in G.*

Proof. The result follows immediately from Propositions 3.3 and 3.4.

PROPOSITION 3.6. *Let G be a plane graph with edge weights. Let $\mathbf{C}$ be an independent collection of cycles of G. If C is a cycle that separates two regions that are not separated by any cycle in $\mathbf{C}$, then $\mathbf{C} \cup \{C\}$ is an independent collection of cycles.*

Proof. The result follows immediately from (3.1) and Proposition 3.4.

PROPOSITION 3.7. *Let G be a plane graph with edge weights. If $\mathbf{C}$ is a nested cycle basis, then for each $C \in \mathbf{C}$ there exists a pair of regions such that C is the only cycle in $\mathbf{C}$ that separates these regions.*

Proof. Let C be an arbitrary cycle in $\mathbf{C}$. We produce a pair of regions separated only by C.

We have shown in Proposition 3.2 that $D_{\mathbf{C}}$ satisfies (a)–(c) in Theorem 2.1. If C is an isolated node in $D_{\mathbf{C}}$, then C is the only cycle in $\mathbf{C}$ that separates the region interior to it from the unbounded region. If C is a root node of $D_{\mathbf{C}}$ (and not isolated), then it is the only cycle in $\mathbf{C}$ separating its child in $\mathbf{F}_{\mathbf{C}}$ from the unbounded region. Otherwise, let C' be the parent of C in $D_{\mathbf{C}}$. If C is a leaf node of $D_{\mathbf{C}}$, then C is the only cycle in $\mathbf{C}$ separating the region interior to it from the child of C' in $\mathbf{F}_{\mathbf{C}}$. If C is not a leaf, then C is the only cycle in $\mathbf{C}$ separating the child of C' in $\mathbf{F}_{\mathbf{C}}$ from the child of C in $\mathbf{F}_{\mathbf{C}}$. □

PROPOSITION 3.8. *Let G be a plane graph with edge weights. Let $\mathbf{C}$ be a nested independent collection of cycles of G. If C is a cycle such that $\mathbf{C} \cup C$ is independent, then C separates some pair of regions not separated by any cycle in $\mathbf{C}$.*

Proof. Consider the tree D_C for the given collection $\mathbf{C}$. $\mathbf{C}$ need not be a basis; hence a node, say C', in $D_{\mathbf{C}}$ may have more than one child in $\mathbf{F}_{\mathbf{C}}$. However, the sum of the children in $\mathbf{F}_{\mathbf{C}}$ of such a cycle C' is given by the sum of C' and the children of C' in $\mathbf{C}$.

Suppose that the cycle C separates no pair of regions that is not already separated by a cycle in $\mathbf{C}$. Then, if it contains a face in $\mathbf{F}_{\mathbf{C}}$, it must contain all siblings of this face in $D_{\mathbf{C}}$. Since C is the sum of the faces it contains, it follows from our remarks above that C can be expressed as a sum of cycles in $\mathbf{C}$. Hence $\mathbf{C} \cup C$ is dependent, and the result follows. □

Proof of (i) *if and only if* (ii). ($\Leftarrow$) Assume that $\mathbf{C}$ satisfies the conditions of (ii) of the theorem. We show that if we apply the greedy algorithm to the set of all cycles in G, then we can produce precisely $\mathbf{C}$. Suppose that by applying the greedy algorithm to the cycles in G, we have chosen $\mathbf{C}' \subset \mathbf{C}$ and we are considering the cycles from G

that we can add in the next iteration of the greedy algorithm. By Proposition 3.8, any candidate cycle must separate two not-yet-separated regions. By Proposition 3.6, any cycle that separates two not-yet-separated regions is independent of $\mathbf{C}'$. Hence, the minimum-weight cycles that separate two not-yet-separated regions are precisely the candidate cycles. However, such a candidate must exist in $\mathbf{C}$, by its definition.

($\Rightarrow$) Assume that $\mathbf{C}$ is an MCB(G). By Proposition 3.5, for every pair of regions in $G, \mathbf{C}$ contains a minimum weight cycle that separates the pair. In addition, by Proposition 3.7, every $C \in \mathbf{C}$ separates a pair of regions that is separated by no other cycle in $\mathbf{C}$. Hence $\mathbf{C}$ also satisfies the minimality requirement of part (ii) of the theorem. This completes the proof. □

4. An MCB algorithm for simple planar graphs. Our main objective in this section is to present an algorithm that finds an MCB of a simple planar graph with nonnegative weights in $O(n^2 \log n)$ time. Before giving the algorithm, we present some useful results on MCBs and four simple algorithms, which serve as subroutines in the main algorithm. One of our first objectives is to prove a slightly different version of Theorem 2.1 (i.e., Theorem 4.7) that will be more useful algorithmically.

Let $G = (V, E)$ be a graph. In the remainder of the paper, we assume that $V = \{v_1, \ldots, v_n\}$. If $V' \subseteq V$, then $\min(V')$ is defined to be the minimum subscript of a node in V'. For any subgraph G' of G, we let $V(G')$ denote the node set of G'. If w is a vector of edge weights for G and P is a path, then we let $w(P)$ denote the sum of the weights of the edges in P and we let $\text{len}(P)$ denote the number of edges in P. We may refer to a path as a *u–v path* if it has endnodes u and v. We use the term *shortest path* to refer to a path that has the minimum weight of all paths with the same endnodes.

PROPOSITION 4.1. *Let $G = (V, E)$ be a simple graph with edge weight vector w. Then, for every pair of nodes, $u, v \in V$, there exists a unique u–v path, call it P, that satisfies exactly one of the following three conditions with respect to any other u–v path P':*

$$w(P) < w(P'), \tag{4.1}$$

$$w(P) = w(P') \quad \textit{and} \quad \text{len}(P) < \text{len}(P'), \tag{4.2}$$

(4.3)

$$w(P) = w(P'), \quad \text{len}(P) = \text{len}(P'), \quad \min(V(P) \setminus V(P')) < \min(V(P') \setminus V(P)).$$

Proof. Consider the set $\mathbf{P}$ of all shortest u–v paths P^* for which $\text{len}(P^*)$ is a minimum. It suffices to show that for every $P_1, P_2 \in \mathbf{P}, V(P_1) \neq V(P_2)$. (It follows immediately that $V(P_1) \setminus V(P_2) \neq \emptyset$ and $V(P_2) \setminus V(P_1) \neq \emptyset$, since P_1 and P_2 have the same number of nodes.)

Assume that $V(P_1) = V(P_2)$. Let us rename the nodes so that P_1 contains the nodes $u_1, \ldots, u_k$, in this order. At some point, P_2 must differ in this sequence from P_1. Let j be the smallest subscript such that u_j is not the jth node in P_2. Consider the path that consists of P_1 from u_1 to u_j plus P_2 from u_j to u_k. This path consists of a shortest path from u_1 to u_j plus a shortest path from u_j to u_k. Hence it is a shortest path from u_1 to u_k. However, it contains fewer edges than P_1 or P_2, which contradicts our choice of $\mathbf{P}$. □

If G is an edge-weighted graph, then we let $P(uv)$ denote the unique u–v path described in Proposition 4.1. We immediately obtain the following simple proposition.

PROPOSITION 4.2. *Let $G = (V, E)$ be a simple edge-weighted graph and let $u, v \in V$. Then*

$$P(uv) \text{ is a shortest } u\text{–}v \text{ path;} \tag{4.4}$$

$$\text{if an } x\text{–}y \text{ path } P \text{ is a subpath of } P(uv)\,, \text{ then } P = P(xy)\,. \tag{4.5}$$

Suppose that w is a vector of edge weights for a graph G. Then a vector of edge weights w' is called a perturbation of w if, for any two-edge subsets, say E_1 and $E_2, w(E_1) < w(E_2)$ implies that $w'(E_1) < w'(E_2)$. Hence, if $\mathbf{C}$ is an MCB(G) under w', then it is an MCB(G) under w.

PROPOSITION 4.3. *If $G = (V, E)$ is a simple graph with edge weight vector w, then there exists a perturbation w' of w such that every path $P(uv)$ under w is the unique shortest u–v path under w'.*

Proof. Let $w(uv)$ denote the original weight of edge uv. Let $\varepsilon, \varepsilon_1, \ldots, \varepsilon_n$ be very small numbers, where $\varepsilon \gg \varepsilon_1 \gg \varepsilon_2 \gg \cdots \gg \varepsilon_n$. We define $w'(v_iv_j) \equiv w(v_iv_j) + \varepsilon - \varepsilon_i - \varepsilon_j$. Since the ε's are small, any shortest path under w' is a shortest path under w. Adding ε to each edge weight insures that condition (4.2) is satisfied, and subtracting the ε_i and ε_j terms insures that condition (4.3) is satisfied. □

The key to making the main algorithm more efficient is to considerably limit the set of candidate cycles. Let C be a cycle of a graph G with edge weights and let $G(C)$ be the subgraph of G with edge set C (and no isolated nodes). Then C is called a *short cycle* if, for every pair of nodes x, y on $G(C)$, a shortest path from x to y in G is contained in $G(C)$.

PROPOSITION 4.4 (see Horton [14]). *Let G be a graph with edge weights. If C is contained in some* MCB(G), *then C is a short cycle.*

If G is a simple edge weighted graph, then we call a cycle C *lexicographically short* (*lex short*) if, for every pair of nodes x, y in $C, P(xy)$ is contained in C. Hence, by Proposition 4.2, if a cycle is lex short, then it is short.

PROPOSITION 4.5. *Let G be a simple graph with edge weights. Then the collection of lex short cycles contains an* MCB(G).

Proof. The proposition follows immediately from the two above propositions.

We next present two useful results concerning the lex short cycles.

PROPOSITION 4.6. *Let G be a simple plane graph with edge weights. Then the collection of lex short cycles is nested.*

Proof. We observe that, if two cycles C and C' are lex short, then $G(C) \cap G(C')$ is a path or empty (otherwise, we contradict the definition of lex short cycles), which is true if and only if C and C' are nested. □

In addition, we immediately obtain the following version of Theorem 2.1.

THEOREM 4.7. *Let G be a simple plane graph with edge weights; let $\mathbf{C}$ be a collection of lex short cycles. Then* (i)–(iii) *of Theorem* 2.1 *are equivalent.*

Proof. This theorem follows immediately from Propositions 4.5 and 4.6. □

In the algorithm for finding an MCB of a plane graph, we generate all the lex short cycles and then extract a subset $\mathbf{C}$ that satisfies (ii) and (iii) of Theorem 2.1. We generate these cycles in a subroutine called Algorithm: Lex Short Cycles. As we next show, this collection of cycles has cardinality $O(n)$ for a simple planar graph, which makes this list of candidates small and gives us our desired complexity. Actually, we show this for any collection of nested cycles.

PROPOSITION 4.8. *Let* $\mathbf{C}$ *be a nested collection of cycles in a simple plane graph* G *and let* f *be the number of (internal) faces of* G. *Then* $|\mathbf{C}| \leq 2f - 1$.

Proof. Assume, without loss of generality, that $\mathbf{C}$ contains all the (internal) faces of G. It follows that $D_{\mathbf{C}}$ is a forest with $f = |\{\text{leaves and isolated nodes}\}|$ and that every node, which is neither a leaf nor isolated, has at least two children. By a straightforward induction on f, we can show that the number of nodes in such a directed graph is less than or equal to $2f - 1$. The result follows. □

COROLLARY 4.9. *Let* G *be a simple plane graph with edge weights. Then*

(4.6) *the cardinality of the collection of lex short cycles of* G *is* $O(n)$.

Proof. The proof follows immediately from Propositions 4.6 and 4.8. □

The first step in generating the lex short cycles is generating the short paths $P(uv)$, which we do in Algorithm: Lex Short Paths. The following notation is used in this algorithm. T_r denotes a partial shortest path tree grown from node r, where, if there is a path from r to u in T_r, then this path is $P(ru)$. For each $u \neq r$ in $T_r, \mathrm{sub}_r(u) \equiv$ the smallest subscript of a node on the path in T_r from u to r, excluding r. $\mathrm{sub}_r(r) \equiv$ the subscript of r. For each $u \neq r$ in $T_r, \mathrm{edge}_r(u) \equiv$ the edge incident with r on the path in T_r from u to r. $\mathrm{edge}_r(r) \equiv \emptyset$. $d(u,v) \equiv$ the weight of a shortest path (i.e., the distance) from u to v. $\mathbf{L}$ denotes an ordered list of the node pairs in nondecreasing order of the distances between the pairs (ties are broken arbitrarily).

The idea of the algorithm is that, at iteration k, the algorithm processes the kth pair on $\mathbf{L}$, say (u,v), by adding u to T_v and adding v to T_u so that $P(uv)$ occurs in both trees.

```
Algorithm: Lex Short Paths
Input: Simple graph $G = (V, E)$ with nonnegative edge weights $w$.
Output: Shortest path trees $T_r$ for each node $r$, where each shortest path from
    $r$ to $u$ in $T_r$ is $P(ru)$.
Step 0: Add $\varepsilon > 0$ to the weight of every edge.
Step 1: Using any shortest path algorithm, construct the list $\mathbf{L}$ as described
    above. Initialize $T_r := r, \mathrm{sub}_r(r) :=$ the subscript of $r, \mathrm{edge}_r(r) := \emptyset$, for each node $r$.
Step 2: For $k = 1, 2, 3, \ldots, |\mathbf{L}|$, do the following:
    Let $(u,v)$ be the $k$th pair on $\mathbf{L}$. Let $u_1, \ldots, u_p$ be the nodes such that (i) $u_i$ is
    incident with $v$ and is contained in $T_u$ and (ii) the path $P(u_iu) \cup u_iv$ has weight
    $d(u,v)$. Similarly, let $v_1, \ldots, v_q$ be those nodes such that (i) $v_i$ is incident with $u$ and
    is contained in $T_v$ and (ii) the path $P(v_iv) \cup v_iu$ has weight $d(u,v)$.
    (a) Identify node pairs $\{u_i, v_j\}$ such that $\mathrm{edge}_u(u_i) = v_ju$ and $\mathrm{edge}_v(v_j) = u_iv$;
        or $u_iv_j =$ the edge $uv$.
    (b) Pick that pair $\{u', v'\}$ for which $\min(\mathrm{sub}(u_i) \cup \mathrm{sub}(v_j))$ is a minimum.
    (c) Add the edges $u'v$ and $v'u$ to $T_u$ and $T_v$, respectively.
    (d) Update the sub( ) and edge( ) functions.
End.
```

PROPOSITION 4.10. *Algorithm: Lex Short Paths works.*

Proof. Note that we need not actually pick a number ε in Step 0, but can simply add this symbol, which represents an arbitrarily small positive number, to the weight of each edge. Under this new weighting, each path has a weight with two terms, one of which is its original weight and one of which is an ε term. Observe that under this new weighting we may assume, as under the original weighting, that additions and comparisons of two numbers can be performed in constant time; thus Step 0 will have no effect upon our complexity arguments. The effect of adding $\varepsilon > 0$ to every edge weight is that if two paths had the same weight but different lengths under the original weights, then the path with fewer edges has less weight under the new weights. Hence $P(uv)$ is the same under both weight vectors. Also, all edge weights become effectively positive, which is important in the remainder of the proof.

The proof proceeds by induction on k in Step 2. Assume inductively that the algorithm has correctly constructed the partial shortest path trees T_r for the first $k-1$ entries on the list **L**. Let us consider the kth node pair (u, v) in Step 2. If there is an edge uv and $d(u, v) =$ the weight of uv, then the algorithm adds uv to both T_u and T_v. So let us assume this is not the case. Consider the path $P(uv)$. Let u_i and v_j be the nodes of $P(uv)$ incident with v and u, respectively. Then, since all edge weights are positive, $d(u_i, u) < d(u, v)$ and $d(v_j, v) < d(u, v)$. Hence, by inductive hypothesis, $P(u_iu)$ and $P(v_jv)$ must be contained in T_u and T_v, respectively. Since $P(u_iu)$ and $P(v_jv)$ are contained in $P(uv)$, the node pair $\{u_i, v_j\}$ is picked in Step 2 (b). Since both T_u and T_v are trees, it is easy to see that all other node pairs identified in Step 2 correspond to u–v paths that are internally node disjoint from $P(uv)$ and each other. Hence Step 2 picks the pair $\{u_i, v_j\}$. □

PROPOSITION 4.11. *Algorithm: Lex Short Paths can be implemented in time $O(n^2 \log n)$ for simple planar graphs.*

Proof. Step 0 can be accomplished in $O(n)$ time (see the discussion in the first paragraph of the proof of Proposition 4.10). Step 1 can be implemented in time $O(n^2 \log n)$, since it takes $O(n^2 \log n)$ time to compute the $d(u, v)$s (e.g., using the planar shortest path algorithm in [18]), of which there are $O(n^2)$, and another $O(n^2 \log n)$ time to sort them to obtain **L**.

In Step 2, $u_1, \ldots, u_p$ and $v_1, \ldots, v_q$ can be found in $O(n)$ time by scanning the neighbors of u and v. The node pairs $\{u_i, v_j\}$ can be found by examining $\text{edge}_u(u_i)$, for $i = 1, \ldots, p$, checking if each such edge contains a node in $v_1, \ldots, v_q$, and then repeating the process by examining $\text{edge}_v(v_j)$ for $j = 1, \ldots, q$. This also requires $O(n)$ time. Hence Step 1 dominates the complexity of this algorithm, and the result follows. □

It will be useful to keep track of another piece of information when Algorithm: Lex Short Paths is run. In particular, we want to know, for each pair of nodes u and v, which edge is in the "middle" of $P(uv)$. Recall that we denote the number of edges in $P(uv)$ by $\text{len}(P(uv))$. First, let $\text{halfnode}_u(v) \equiv$ the node w on $P(uv)$ such that $\text{len}(P(uw)) = \text{len}(P(wv))$, when $\text{len}(P(uv))$ is even; $\text{halfnode}_u(v) \equiv$ the node w on $P(uv)$ such that $\text{len}(P(uw)) = \text{len}(P(wv)) - 1$, when $\text{len}(P(uv))$ is odd. Suppose that we are in Step 2 of the algorithm and considering a pair of nodes u and v. Then it is easy to see that $\text{halfnode}_u(v)$ and $\text{halfnode}_v(u)$ can be computed in constant time from $\text{halfnode}_u(u')$ and $\text{halfnode}_v(v')$, where u' and v' are defined as in Step 2.

Finally, let u and v be two nodes and assume, without loss of generality, that the subscript of u is less than the subscript of v. Then $\text{halfedge}(uv) \equiv$ the edge between $\text{halfnode}_u(v)$ and $\text{halfnode}_v(u)$, when $\text{len}(P(uv))$ is odd; $\text{halfedge}(uv) \equiv$ the edge in $P(uv)$ incident with $\text{halfnode}_u(v)(= \text{halfnode}_v(u))$ and closest to u, when $\text{len}(P(uv))$ is even. Again, we can compute, in constant time, $\text{halfedge}(uv)$ for each uv when it is considered in Step 2 of the algorithm.

The first step in the algorithm for generating the lex short cycles is to generate a superset of these cycles described in the following elementary proposition, which follows immediately from a simple proposition of Horton [14]. Let us call a cycle C in a simple edge-weighted graph G a *lex edge-short cycle with respect to a node x* if, for a node x in $G(C)$, there exists an edge yz in $G(C)$ such that $G(C) = P(xy) \cup$

$P(xz) \cup yz$. Note that, by their definition, $P(xy)$ and $P(xz)$ share only the node x. As a shorthand, a lex edge-short cycle C as just described is sometimes referred to as $\{x, yz\}$.

Due to the following proposition, we can use the lex edge-short cycles as a first step in obtaining the lex short cycles. This idea is due to Horton [14].

PROPOSITION 4.12 (see Horton [14]). *Let G be a simple graph with edge weights. Then C is a lex short cycle if and only if C is a lex edge-short cycle with respect to every node in C.*

Proof. This result follows immediately from Theorem 4 in [14].

The following proposition proves useful. We leave its proof to the reader.

PROPOSITION 4.13. *Let G be a simple edge-weighted graph. If C is a lex edge-short cycle with respect to a node x, then there exists a unique edge yz in $G(C)$ such that $G(C) = P(xy) \cup P(xz) \cup yz$, where $P(xy), P(xz)$, and yz are edge disjoint.*

OBSERVATION 4.14. *In a simple planar graph, there are $O(n^2)$ lex edge-short cycles, since there are $O(n^2)$ node-edge pairs.*

To extract the lex short cycles from the list of lex edge-short cycles, we use a simple data structure defined in and constructed by the following algorithm. A *directed tree rooted at r*, say $T = (V, A)$, is a directed tree whose arcs are oriented away from the node r. As before, if $xy \in A$, we say y is a *child* of x; also, a *leaf* of T is a node within degree 1. If there exists a directed path from node a to node b, then we say b is a *descendent* of a.

Algorithm: Interval

Input: A tree $T = (V, A)$ with root node r.

Output: A closed interval for each node in V.

Step 1: Orient the edges of T to obtain a directed tree D rooted at r.

Step 2: Perform a depth first search on the nodes of D beginning at r. Label the leaves of D in the order they are scanned with 1, 2, 3, ...

Step 3: For each $v \in V$, set $\text{interval}_r(v) := [a, b]$, where a is the smallest label of a leaf that is a descendent of v and b is the largest label of a leaf that is a descendent of v.

End.

We easily obtain the following proposition.

PROPOSITION 4.15. *Suppose that Algorithm: Interval is applied to a tree $T = (V, A)$ with root node r. Let $x, y \in V$. Then x is on the path of T from r to y if and only if* $\text{interval}_r(y) \subseteq \text{interval}_r(x)$.

Hence, if we have the above data structure for a particular tree T_r, then we can answer, in constant time, the question "Is y a descendent of x in T_r?"

PROPOSITION 4.16. *For a tree $T = (V, A)$ with root node r, Algorithm: Interval can be implemented in $O(n)$ time.*

Proof. Steps 1 and 2 can be performed in O(n) time. Step 3 can also be performed in $O(n)$ time as follows. For each node v in T, we determine a and b, where $\text{interval}_r(v) = [a, b]$, as follows. When v is first scanned in the depth first search, note this, and let a be the next number attached to a leaf. When v is last scanned, let b be the last number attached to a leaf. □

In the algorithm for generating the lex short cycles, we use the following algorithm. It takes as input a lex edge-short cycle and tests an obvious necessary condition for the cycle to be a lex short cycle. When the cycle passes this test, it also returns a "canonical" representation for the candidate cycle.

Algorithm: Lex Short Cycle Test

Input: A simple graph $G = (V, E)$ with edge weights and a lex edge-short cycle $C = \{v, xy\}$.

Output: *Yes*, if C is lex edge-short with respect to z, where z is the node of C with smallest subscript; otherwise output *no*. If yes, then also output $\{z, ab\}$, where $\{z, ab\} = C$.

Step 1: Find the node z in C with the smallest subscript.

Step 2: If $z = v$, then output *yes* and $\{v, xy\}$. If $z \neq v$, then, for each edge ab in C, check if $P(za) \cup P(zb) \cup P(ab) = G(C)$. If yes for some ab, then output *yes* and $\{z, ab\}$. Otherwise, output *no*.

End.

The validity of the algorithm follows immediately from the above propositions. The obvious implementation takes linear time, but with a little care we can implement it faster as follows.

PROPOSITION 4.17. *If* $\text{interval}_u(v)$ *and* $\text{halfedge}_u(v)$ *for every pair of nodes* u *and if* v *of* G *is included in the input to Algorithm: Lex Short Cycle Test, then it can be implemented in time* $O(\log n)$.

Proof. Let $C = \{v, xy\}$ and z be as in the algorithm. Suppose that $z \neq v$ and, without loss of generality, suppose that z occurs on $P(vx)$. Since $P(zv)$ and $P(zx)$ are subpaths of $P(vx)$ (by Proposition 4.2), we need only check if $ab = xy$ or ab occurs on $P(vy)$. Suppose that $ab = xy$. Then we need only check if v occurs on $P(zy)$. (This can be done in constant time by checking if $\text{interval}_z(y) \subseteq \text{interval}_z(v)$.) If the answer to this is yes, then (by Proposition 4.2) $C = P(za) \cup P(zb) \cup ab$.

Suppose that ab occurs on $P(vy)$. We perform a binary search for this edge as follows. Let $\text{halfedge}(vy) = a_1b_1$. Suppose that a_1 is closer to y than v (we can determine this in constant time by checking, for example, if a_1 is contained in $P(b_1y)$). Finally, we must check (i) if x and y occur on $P(za_1)$ and if $P(xy) = xy$; and (ii) if v occurs on $P(zb_1)$. If both are true, then $a_1b_1 = ab$, and we are done. If neither is true, then C is not lex short. If (i) is true and (ii) is false, then we can restrict our search for ab to $P(a_1y)$. If (i) is false and (ii) is true, then we can restrict our search for ab to $P(b_1v)$. Each of these checks can be performed in constant time (using the data structure defined in Algorithm: Interval and discussed in Proposition 4.15). We then apply the same procedure to this subpath of $P(vy)$. The total work is bounded by $O(\log n)$. □

In the algorithm for generating the lex short cycles when the weights are nonnegative, we first generate the lex edge-short cycles (Horton in [14] generates the same collection but starting from an arbitrary collection of shortest paths). However, there are $O(n^2)$ such cycles; hence it requires $O(n^3)$ time to write their incidence vectors. Since our algorithm for finding the MCB runs in $O(n^2 \log n)$ time, we do not write out the entire list (as Horton does). The idea is to find these cycles without explicitly writing them out. This is what we do in Step 3 of the following algorithm. Then, in Steps 4 and 5, we identify and explicitly write out just the lex short cycles.

Algorithm: Lex Short Cycles

Input: Simple planar graph $G = (V, E)$ with nonnegative edge weights w.

Output: The incidence vectors of the collection $\mathbf{L}$ of lex short cycles.

Step 1: Set $\mathbf{L}, \mathbf{L}'$, and $\mathbf{L}'' := \emptyset$.

Step 2: Apply Algorithm: Lex Short Paths. Let $T_r = (V, E_r)$ be the shortest path tree rooted at r.

Step 3: Generate the lex edge-short cycles:

For each $r \in V$, do the following:

For every $uv \in E \setminus E_r$, let $C(r, uv)$ refer to the unique cycle in $uv \cup E_r$; if $C(r, uv)$ contains r, then add $(\{r, uv\}, \text{len}(\{r, uv\}))$ to $\mathbf{L}'$, where $\text{len}(r, uv)$ denotes the number of edges in $\{r, uv\}$.

Step 4: For each $(\{r, uv\}, \text{len}(r, uv))$ in $\mathbf{L}'$, apply Algorithm: Lex Short Cycle Test to $\{r, uv\}$. If the output is yes, then replace $(\{r, uv\}, \text{len}(r, uv))$ with $(\{w, xy\}, \text{len}(w, xy))$ on $\mathbf{L}'$, where xy are ordered so that x has smaller subscript than y.

Step 5: Sort the members of $\mathbf{L}'$ lexicographically (say, from left to right in ascending order on the subscripts of the three nodes in the node and edge that define it). For every $(\{w, xy\}, \text{len}(w, xy))$ on $\mathbf{L}'$, if it occurs $\text{len}(w, xy)$ times in $\mathbf{L}'$, then add it (once) to $\mathbf{L}''$.

Step 6: Write out in $\mathbf{L}$ the incidence vectors of the cycles on $\mathbf{L}''$.

End.

PROPOSITION 4.18. *Algorithm: Lex Short Cycles works.*

Proof. Clearly, we generate, in Step 3, the lex edge-short cycles. By Proposition 4.12, each short cycle is generated in Step 3 once for each node in the cycle. Also, any cycle that is generated this number of times must be a short cycle. Hence the short cycles are characterized by how often they appear in $\mathbf{L}'$. This is what we test for in Step 5, since in Step 4 we find a unique representation for each short cycle in $\mathbf{L}'$. Hence the algorithm works. □

PROPOSITION 4.19. *Algorithm: Lex Short Cycles can be implemented in $O(n^2 \log n)$ time for simple planar graphs.*

Proof. By Proposition 4.11, Step 2 requires $O(n^2 \log n)$ time. Let $\text{len}_r(v)$ be the number of edges on the path in T_r from r to v. Then, for every $r, v \in V$, we can also record $\text{len}_r(v)$ in Step 2. Observe that in Step 3, $C(r, uv)$ contains r if and only if $\text{edge}_r(u) \neq \text{edge}_r(v)$. Hence we can check if $C(r, uv)$ contains r in constant time, and, if the answer is yes, we have $\text{len}(r, uv) = \text{len}_r(u) + \text{len}_r(v) + 1$. Since G has $O(n)$ edges, the complexity of Step 3 is $O(n^2)$. The key idea in this step is not explicitly to write out the incidence vectors of the cycles generated, which would require $O(n^3)$ time.

By Proposition 4.17, Step 4 can be implemented in time $O(n^2 \log n)$ since $\mathbf{L}'$ has length $\text{O}(n^2)$.

Step 5 is dominated by the sorting, which requires $O(n^2 \log n)$ time.

By Corollary 4.9, $O(n)$ cycles are added to $\mathbf{L}$ in Step 6, each of which requires $O(n)$ time to write out. Hence Step 5 requires $O(n^2)$ time.

So the overall complexity is dominated by Steps 2, 4, and 5, which proves the result. □

Algorithm: MCB

Input: A simple planar graph G with nonnegative edge weights.

Output: An MCB(G).

Step 1: Embed G in the plane and find $\mathbf{F_C}$, the list of its faces.

Step 2: Apply Algorithm: Lex Short Cycles to G to get a list of candidate cycles $\mathbf{L}$ (i.e., $\mathbf{L}$ contains an MCB(G)) and sort $\mathbf{L}$ in ascending order by cycle weights.

Step 3: Set $\mathbf{C} := \emptyset$ and $D_{\mathbf{C}} := (\mathbf{F_C}, A)$, where $A := \emptyset$.

Step 4: If there exist $C_1, C_2 \in \mathbf{F_C}$ with the same parent in $D_{\mathbf{C}}$, then set $R_1 :=$ interior(C_1) and $R_2 :=$ interior(C_2); otherwise, let C_1 be an isolated node of $D_{\mathbf{C}}$ in $\mathbf{F_C}$ and set $R_1 :=$ interior(C_1) and $R_2 :=$ exterior region.

Step 5: Find a minimum weight cycle $C \in \mathbf{L}$ that separates R_1 and R_2. Set $\mathbf{C} := \{\mathbf{C} \cup C\}$.

Step 6: If $|C| <$ dimension of G, then update $D_{\mathbf{C}}$ with a node for C and go to Step 4. Otherwise, $\mathbf{C} = \text{MCB}(G)$.

End.

OBSERVATION 4.20. *If we wish to drop the condition that G be simple, then we can apply Algorithm: Reduction to G, which appears in the last section. This algorithm outputs a simple planar graph G' plus a collection of cycles $\mathbf{C}'$ such that the output of Algorithm:* MCB *applied to G' plus the cycles in $\mathbf{C}'$ is an* $\text{MCB}(G)$. *Note that if we are solving the* APMC *problem on a simple plane graph G, then G^d may not be simple. However, the number of edges in G^d is still $O(n)$; hence first applying Algorithm: Reduction to G^d and then applying Algorithm:* MCB *still requires $O(n^2 \log n)$ time.*

PROPOSITION 4.21. *Algorithm:* MCB *works.*

Proof. This follows immediately from equivalent conditions (ii) and (iii) in Theorem 4.7. □

PROPOSITION 4.22. *Algorithm:* MCB *can be implemented in time $O(n^2 \log n)$.*

Proof. Let us observe that it is sufficient that $\mathbf{F}_{\mathbf{C}}$ consists of pointers to the faces of G.

Step 1 takes $O(n)$ time (see, e.g., [2]). Step 2 is dominated by Algorithm: Lex Short Cycles and the sort, each of which takes $O(n^2 \log n)$ time.

Step 3 takes $O(n)$ time. Step 4 requires a breadth-first scanning of D_C. By Proposition 4.8, $D_{\mathbf{C}}$ has $O(n)$ nodes; hence this step requires $O(n)$ time.

Step 5 can be performed in $O(n)$ time as follows. First, at the end of Step 2, it is easy to construct, in $O(n^2)$ time, the incidence matrix of cycles in $\mathbf{L}$ versus interior regions (i.e., $a_{ij} = 1$ if and only if cycle i contains region j). Then, in Step 5, by scanning the two columns of this matrix that correspond to R_1 and R_2, the minimum weight separating cycle can be found in $O(n)$ time.

We finally consider the complexity of updating $D_{\mathbf{C}}$ in Step 6. Let C be a new cycle found in Step 5. From the output of Step 2, we know the regions/faces interior to C. These faces each have a corresponding leaf node of $D_{\mathbf{C}}$. Let V' be this set of leaves. If C_1 and C_2 have the same parent P in $D_{\mathbf{C}}$, then C becomes a child of P, and all branches of the tree that descended from P to a node in V' now descend from C. If C_1 is isolated, then, again, all maximal branches of $D_{\mathbf{C}}$ that descend to a node in V' now descend from C. Hence Step 6 can be performed in $O(n)$ time.

So Step 2 dominates the complexity of the algorithm, and the result follows. □

5. Extension to general graphs. The following algorithm allows us to reduce the MCB problem on general graphs to the MCB problem on simple graphs. Hence, with this algorithm and Algorithm: MCB, we can solve the MCB problem on general planar graphs in $O(n^2 \log n + m)$ time (see Observation 4.20).

Algorithm: MCB Reduction

Input: A graph $G = (V, E)$ with nonnegative edge weights w.

Output: A subgraph $G' = (V', E')$ of G with nonnegative edge weights and a collection of cycles $\mathbf{C}'$ such that G' is simple and an $\text{MCB}(G')$ together with $\mathbf{C}'$ is an $\text{MCB}(G)$.

Step 1: Let $\mathbf{C}' := \emptyset, G' = (V', E')$ where $V' := V$ and $E' := E$.

Step 2: Let $L \subseteq E$ be the loops of G. Set $\mathbf{C}' := \mathbf{C}' \cup L$. Set $E' := E' \setminus L$.

Step 3: For each maximal collection in G of parallel edges $e_1, \ldots, e_p, p \geq 2$, with the same endnodes, do the following:

(i) Rename the edges so that $w(e_1) \leq w(e_i), i \geq 2$.

(ii) Set $E' := E' \setminus \{e_i \colon i \geq 2\}$.

Step 4: For each pair of nodes u and v of V, find Path(u, v), a shortest u–v path in G'.

Step 5: For each maximal collection in G of parallel edges $e_1, \ldots, e_p, p \geq 2$, with the same endnodes, say u and v, do the following:

$$\text{Set } \mathbf{C}' := \mathbf{C}' \cup \{E(\text{Path}(u, v)) \cup e_i\}, \text{ for each } i \geq 2.$$

End.

PROPOSITION 5.1. *Algorithm: Reduction works.*

Proof. Let w' be obtained from w as follows: For each maximal collection in G of parallel edges $e_1, \ldots, e_p, p \geq 2$, add $\varepsilon > 0$ to $w(e_i), i \geq 2$. Clearly, an MCB(G) under w' is also an MCB(G) under w.

From the definition of cycle basis, it immediately follows that each edge of G is contained in at least one cycle of any MCB. Let $e = uv$ be an arbitrary edge whose weight has been altered in constructing w'. By our choice of w', e is not a shortest path. Hence, by Proposition 4.4, each cycle in an MCB that contains e must consist of a shortest u–v path in G' together with e. Suppose that there are two such cycles, say C_1 and C_2, through e in some MCB(G) under w'. Call it $\mathbf{C}$. Let $C = C_1 + C_2$. By our above observation, $w'(C) < w'(C_j), j = 1, 2$. By our definition of C, either C_1 or C_2 can be replaced by C in $\mathbf{C}$ to yield a new basis, say $\mathbf{C}'$, for the cycle space with less weight than $\mathbf{C}$. (If C is not a cycle, then it is the sum of edge disjoint cycles and therefore it can be replaced in $\mathbf{C}'$ by one of these cycles to yield a cycle basis of the same weight. For a further discussion of this idea, see [14]. Hence we may assume that $\mathbf{C}'$ contains only cycles.) However, this contradicts our choice of $\mathbf{C}$. Hence any MCB(G) under w' must contain exactly one cycle through each edge whose weight we have altered; this cycle must be of the form we identify in Step 5. The result follows. □

PROPOSITION 5.2. *If G is planar, then Algorithm: Reduction can be implemented in time $O(n^2 \log n + m)$.*

Proof. Steps 2 and 3 each require $O(m)$ time. Step 4 requires $O(n^2 \log n)$ time, using any planar shortest path algorithm (e.g., see [18]). Note that writing out the cycles in Step 5 may take $O(mn)$ time if there are more than $O(n)$ parallel edges and if the shortest paths in these cycles are O(n) in length. However, there are $O(n)$ shortest paths in G' between endnodes of edges, and, for parallel edges, a pointer may be used to the corresponding shortest path, rather than writing the cycle out completely. Hence, the collection $\mathbf{C}'$ may be compactly written out in $O(m)$ time. So the complexity of the algorithm is $O(n^2 \log n + m)$. □

Acknowledgments. We thank several anonymous referees for their very helpful suggestions. In particular, the proof of Proposition 4.8 is due to a referee.

REFERENCES

[1] A. C. CASSELL, J. C. HENDERSON, AND K. RAMACHANDRAN, *Cycle bases of minimal measure for the structural analysis of skeletal structures by the flexibility method*, in Proc. Royal Soc. of London Ser. A, 350 (1976), pp. 61–70.

[2] N. CHIBA, T. NISHIZEKI, S. ABE, AND T. OZAWA, *A linear algorithm for embedding planar graphs using PQ-trees*, J. Comput. Systems Sci., 30 (1985), pp. 54–76.

[3] L. O. CHUA AND L. CHEN, *On optimally sparse cycle and coboundary basis for a linear graph*, IEEE Trans. Circuit Theory, CT-20 (1973), pp. 495–503.

[4] D. W. CRIBB, R. D. RINGEISEN, AND D. R. SHIER, *On cycle bases of a graph*, Congr. Numer., 32 (1981), pp. 221–229.

[5] D. Cvetkovic, I. Gutman, and N. Trinajstic, *Graph theory and molecular orbitals* VII: *The role of resonance structures*, J. Chemical Physics, 61 (1974), pp. 2700–2706.

[6] N. Deo, G. M. Prabhu, and M. S. Krishnamoorthy, *Algorithms for generating fundamental cycles in a graph*, ACM Trans. Math. Software, 8 (1982), pp. 26–42.

[7] E. Dijkstra, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[8] E. T. Dixon and S. E. Goodman, *An algorithm for the longest cycle problem*, Networks, 6 (1976), pp. 139–344.

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

[10] G. N. Frederickson, *Fast algorithms for shortest paths in planar graphs, with applications*, SIAM J. Computing, 16 (1987), pp. 1004–1022.

[11] R. Gomory and T. C. Hu, *Multi-terminal network flows*, J. SIAM, 9 (1961), pp. 551–570.

[12] R. Hassin and D. B. Johnson, *An* $O(n \log^2 n)$ *algorithm for maximum flow in undirected planar networks*, SIAM J. Comput., 14 (1985), pp. 612–624.

[13] R. Hassin, *Maximum flow in* (s, t) *planar networks*, Inform. Process Lett., 13 (1981), p. 107.

[14] J. D. Horton, *A polynomial time algorithm to find the shortest cycle basis of a graph*, SIAM J. Comput., 16 (1987), pp. 358–366.

[15] T. C. Hu, *Integer Programming and Network Flows*, Addison–Wesley, Reading, MA, 1969.

[16] ———, *Optimum communication spanning trees*, SIAM J. Comput., 3 (1974), pp. 188–195.

[17] A. Itai and Y. Shiloach, *Maximum flow in planar networks*, SIAM J. Comput., 8 (1979), pp. 135–150.

[18] D. B. Johnson, *Efficient algorithms for shortest paths in sparse networks*, J. Assoc. Comput. Mach., 24 (1977), pp. 1–13.

[19] D. B. Johnson and S. M. Venkatesan, *Using divide and conquer to find flows in directed planar networks in* $O(n^{3/2} \log n)$ *time*, in Proc. Twentieth Annual Allerton Conference on Communication, Control, and Computing, Univ. of Illinois, Urbana, IL, Oct. 1982, pp. 898–905.

[20] A. Kaveh, *An efficient program for generating subminimal cycle bases for the flexibility of structures*, Comm. Applied Numer. Meth., 2 (1986), pp. 339–344.

[21] G. Kirchhoff, *Uber die Auflosung der Gleichungen, auf welche man bei der Untersuchungen der Linearen Verteilung Galvanisher Strome Gefuhrt wird*, Poggendorf Ann. Physik, 72 (1847), pp. 497–508. (English transl. in Trans. Inst. Radio Engrs., CT-5 (1958), pp. 4–7.)

[22] D. E. Knuth, *The Art of Computer Programming*, Vol. 1, Addison–Wesley, Reading, MA, 1968.

[23] E. Lawler, *Combinatorial Optimization*, Holt, Rinehart and Winston, New York, 1976.

[24] S. MacLane, *A structural characterization of planar combinatorial graphs*, Duke Math. J., 3 (1937), pp. 340–472.

[25] P. Matei and N. Deo, *On algorithms for enumerating all circuits of a graph*, SIAM J. Comput., 5 (1976), pp. 90–99.

[26] M. Randic, *Resonance energy of very large benzenoid hydrocarbons*, Internat. J. Quantum Chemistry, XVII (1980), pp. 549–586.

[27] J. H. Reif, *Minimum s–t cut of a planar undirected network in* $O(n \log^2(n))$ *time*, SIAM J. Comput., 12 (1983), pp. 71–81.

[28] Y. Shiloach, *A multi-terminal minimum cut algorithm for planar graphs*, SIAM J. Comput., 9 (1980), pp. 219–225.

[29] N. Trinajstic, *Chemical Graph Theory*, CRC Press, Boca Raton, FL, Vol. 2, 1983.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 419–436, August 1994

007

RANDOM SET PARTITIONS*

WILLIAM M. Y. GOH[†] AND ERIC SCHMUTZ[†]

Abstract. For random partitions of $[n]$, let L_n and R_n, respectively, denote the maximum block size and its multiplicity. The average multiplicity is $E(R_n) = H(\{m_n\}) + o(1)$ as $n \to \infty$, where H is an explicitly given analytic function and $\{m_n\}$ is the fractional part of a certain implicitly defined root. The cumulative distribution function of L_n also depends on $\{m_n\}$. The sequence $\langle\{m_n\}\rangle_{n=1}^{\infty}$ is dense in $(0,1)$. This establishes both the nonexistence of a limit distribution for L_n and the nonexistence of a limiting value for $E(R_n)$.

Key words. random set partitions, blocks

AMS subject classifications. 05A18, 11B75, 60C05

1. Introduction. Several authors have studied random set partitions. We do not review this literature but instead refer the reader to the references ([1]–[4], [6], [10]–[12], [15], [16]). Here we study the maximum block size of random set partitions.

To state the results, we need some notation. If π is a partition of $[n]$, let $L_n(\pi)$ be the cardinality of the largest block of π. Let $u = u_n$ be the positive solution to $ze^z = n$ and let $m_n = \mathrm{e}u - \log\sqrt{u}$. Finally, let $\mu_n = m_n - \log((\mathrm{e}-1)\sqrt{2\pi \mathrm{e}})$. The asymptotic distribution of the maximum block size was determined by Sachkov [16].

THEOREM 1 (Sachkov). *As $n \to \infty$,*

$$\operatorname{Prob}(L_n \le \mu_n + x) = \exp(-\exp(-x + \{\mu_n + x\}))(1 + o(1))$$

uniformly for $|x| \le \log(\sqrt{u})$.

Sachkov did not state his theorem correctly. Nevertheless, he did prove Theorem 1.

The condition $|x| \le \log(\sqrt{u})$ is too stringent for our purposes. We need the following two bounds for the tails of the distribution.

THEOREM 2. *Let a be a fixed real number and let $x = -a \log\log n$. Then, as $n \to \infty$,*

$$\operatorname{Prob}(L_n \le \mu_n + x) = \exp\left(-(\log^a n)\mathrm{e}^{\{\mu_n + x\} + o(1)}\right)(1 + o(1)).$$

THEOREM 3. *Let $b > 0$ be a fixed positive constant and let $x = b \log n$. Then, as $n \to \infty$,*

$$\operatorname{Prob}(L_n \le \mu_n + x) \\ = \exp\left(-\frac{(\mathrm{e}+b)^{c_0 + \{\mu_n + x\}}}{\sqrt{2\pi}\,(b+\mathrm{e})\,(b+\mathrm{e}-1)} \frac{(\log n)^{(\mathrm{e}+1/2)\log(1+b/\mathrm{e}) - b}}{n^{(b+\mathrm{e})\log(1+b/\mathrm{e})}}\right)(1 + o(1)),$$

where $c_0 = \log((\mathrm{e}-1)\sqrt{2\pi \mathrm{e}})$.

* Received by the editors August 27, 1990; accepted for publication (in revised form) June 3, 1993. This research was supported by National Science Foundation grant DMS-8901610.

[†] Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104.

Both these results are needed to determine the average multiplicity of the largest block size.

Let $R_n(\pi)$ be the number of blocks of size $L_n(\pi)$ that π has. Let

$$H(y) = \frac{1}{\sqrt{2\pi e}} e^y \sum_{\xi=-\infty}^{\infty} e^{-\xi} \exp\left(\frac{-1}{\sqrt{2\pi e}\,(e-1)} e^{y-\xi}\right).$$

Remark. H is analytic on a horizontal strip containing the real axis. It is also clear that H is periodic with period 1.

The average multiplicity is given by the following theorem.

THEOREM 4. $E(R_n) = H(\{m_n\}) + o(1)$.

Because of Theorems 1 and 4, we wanted to know more about the sequence of fractional parts. This led to the following result.

THEOREM 5. *The sequence of fractional parts* $\langle\{m_n\}\rangle_{n=1}^{\infty}$ *is dense in* $(0,1)$.

Some interesting corollaries are obtained when Theorem 5 is coupled with Theorems 1 and 4.

COROLLARY 1. *For any fixed* x,

$$\limsup_{n\to\infty} \operatorname{Prob}(L_n \le \mu_n + x) = \exp(-e^{-x}),$$
$$\liminf_{n\to\infty} \operatorname{Prob}(L_n \le \mu_n + x) = \exp(-e^{-x+1}).$$

COROLLARY 2. *We have*

$$\limsup_{n\to\infty} E(R_n) = \max_{0\le y\le 1} H(y) = 1.719398\cdots,$$
$$\liminf_{n\to\infty} E(R_n) = \min_{0\le y\le 1} H(y) = 1.717164\cdots.$$

2. Preliminary estimates.

Both Theorems 2 and 3 are proved by coupling the saddle point method with Szegö's approximations for partial sums of the exponential series. This section contains the proof of Theorem 2. First, we use Evgrafov's formulation of the saddle point method to derive a not-yet-explicit asymptotic formula for $\operatorname{Prob}(L_n \le \mu_n + x)$.

Let $\Psi(n,m)$ be the number of partitions of the set $[n]$ with block sizes that are all less than or equal to m. By the exponential formula [18],

$$\sum_n \frac{\Psi(n,m)}{n!} z^n = \frac{1}{e}\exp(S_m(z)), \tag{1}$$

where $S_m(z) = \sum_{k=0}^{m} z^k/k!$, the mth partial sum of the exponential series. Given a fixed real number $a > 0$, let $x = -a \log\log n$, and let

$$M = \mu_n - a\log\log n = eu - \left(a + \frac{1}{2}\right)\log\log n - c_0 + o(1), \quad \text{where}$$
$$c_0 = \log\left((e-1)\sqrt{2\pi e}\right).$$

The case when $a \le 0$ can be treated in a similar fashion.

By Cauchy's theorem,

$$\text{Prob}\left(L_n \le \mu_n + x\right) = \frac{\Psi\left(n, \lfloor M \rfloor\right)}{B\left(n\right)} = \frac{n!}{2\pi \mathrm{e} \mathrm{i} B\left(n\right)} \oint_C \frac{\exp\left(S_{\lfloor M \rfloor}\left(z\right)\right)}{z^{n+1}}\,dz, \tag{2}$$

where $\lfloor M \rfloor$ is the greatest integer $\le M$; i.e., $\lfloor M \rfloor = M - \{M\}, C$ is a contour encircling the origin, and $B(n)$ is the nth Bell number.

The integral in (2) can be estimated by the saddle point method. Let

$$h\left(n, t\right) = S_{\lfloor M \rfloor}\left(t\right) - n \log t,$$

and define ρ to be the positive solution to

$$\frac{\partial h\left(n, t\right)}{\partial t} = 0. \tag{3}$$

By Theorem 4 of [7, pp. 21, 46][1],

$$\frac{1}{2\pi \mathrm{e} \mathrm{i}} \oint_C \frac{\exp\left(S_{\lfloor M \rfloor}\left(z\right)\right)}{z^{n+1}}\,dz \sim \frac{\exp\left(h\left(n, \rho\right)\right)}{\rho\sqrt{2\pi \frac{\partial^2 h(n,\rho)}{\partial t^2}}} \quad \text{as}\, n \to \infty. \tag{4}$$

We omit the technical details (see the Appendix). Using (3), we obtain

$$h\left(n, \rho\right) = \frac{n}{\rho} - 1 + \frac{\rho^{\lfloor M \rfloor}}{\lfloor M \rfloor!} - n \log \rho. \tag{5}$$

Similarly,

$$\frac{\partial^2 h\left(n, \rho\right)}{\partial t^2} = \frac{n}{\rho} - \frac{\rho^{\lfloor M \rfloor - 1}}{\left(\lfloor M \rfloor - 1\right)!} + \frac{n}{\rho^2}. \tag{6}$$

Putting (5) and (6) into (4), we obtain

$$\frac{n!}{B\left(n\right) 2\pi \mathrm{e} \mathrm{i}} \oint_C \frac{\exp\left(S_{\lfloor M \rfloor}\left(z\right)\right)}{z^{n+1}}\,dz \sim \frac{n! \exp\left(\frac{n}{\rho} - 1 + \frac{\rho^{\lfloor M \rfloor}}{\lfloor M \rfloor!}\right)}{B\left(n\right) \rho^{n+1} \sqrt{2\pi \left(\frac{n}{\rho} - \frac{\rho^{\lfloor M \rfloor - 1}}{\left(\lfloor M \rfloor - 1\right)!} + \frac{n}{\rho^2}\right)}}. \tag{7}$$

It is well known [5] that

$$\frac{n!}{B(n)} \sim \mathrm{e}\sqrt{2\pi}\, \exp\left(-\mathrm{e}^u + u\mathrm{e}^u \log u + u/2 + \log u\right).$$

Putting this into the right side of (7), we obtain

$$\text{Prob}\left(L_n \le M\right) \sim \exp\left(\frac{n}{\rho} - \frac{n}{u} + n \log u - n \log \rho + \frac{\rho^{\lfloor M \rfloor}}{\lfloor M \rfloor!}\right) \cdot \Delta, \tag{8}$$

[1] Beware of typographical errors.

where $\Delta = \exp(u/2 + \log u - \log\rho - \frac{1}{2}\log(n/\rho - \rho^{\lfloor M\rfloor - 1}/(\lfloor M\rfloor - 1)! + n/\rho^2))$.

3. Szegö's approximations. The right side of (8) contains the quantities ρ and u. It is well known [5] that

$$u = \log n - \log\log n + o(1)\,. \tag{9}$$

We need similar estimates for ρ.

PROPOSITION 1. *We have*

$$\rho = u + \frac{\mathrm{e}\log^{a+1} n}{n}\exp\left(\{\mu_n + x\} + o(1)\right).$$

Proof. First, we prove a weaker result,

$$\rho - u = O\left(\frac{\log^{a+1} n}{n}\right). \tag{10}$$

Let $f(t) = tS_{\lfloor M\rfloor -1}(t)$, where $M = \mu_n - a\log\log n$. By the mean value theorem, there is a $\xi \in (u, \rho)$ such that

$$\rho - u = \frac{f(\rho) - f(u)}{f'(\xi)}\,. \tag{11}$$

The tool that enables us to estimate the right side of (11) is the following approximation theorem of Szegö [17].

THEOREM 6 (Szegö). *As* $m \to \infty$,

$$S_m(mw) = \mathrm{e}^{mw}\left(1 - \frac{1}{\sqrt{2\pi m}}\left(\frac{w}{1-w}\right)\left(w\mathrm{e}^{1-w}\right)^m\left(1 + O(1/m)\right)\right),$$

uniformly for $|w| \le \frac{1}{2}$.

Recall that $u\mathrm{e}^u = n$. By taking $w = u/(\lfloor M\rfloor - 1)$ and $m = \lfloor M\rfloor - 1$ in Theorem 6, we obtain

$$\begin{aligned} f(u) = n\Bigg(1 &- \frac{1}{\sqrt{2\pi(\lfloor M\rfloor - 1)}}\left(\frac{u}{\lfloor M\rfloor - 1 - u}\right) \\ &\times\left(\frac{u}{\lfloor M\rfloor - 1}\mathrm{e}^{1-u/(\lfloor M\rfloor - 1)}\right)^{\lfloor M\rfloor - 1}\left(1 + O(1/M)\right)\Bigg). \end{aligned} \tag{12}$$

From the definition of $\rho, f(\rho) = n$. To estimate (11), we still need a lower bound for $f'(\xi)$.

Note that

$$f'(t) = S_{\lfloor M\rfloor -1}(t) + xS_{\lfloor M\rfloor -2}(t)\,. \tag{13}$$

For each $m, S_m(t)$ is an increasing function of t. Hence

$$\begin{aligned} f'(\xi) &\ge S_{\lfloor M\rfloor -1}(u) + uS_{\lfloor M\rfloor -2}(u) \\ &= (1+u)S_{\lfloor M\rfloor -1}(u) - u^{\lfloor M\rfloor}/(\lfloor M\rfloor - 1)!. \end{aligned} \tag{14}$$

It is easy to check, using (9) and Stirling's formula, that

$$u^{\lfloor M\rfloor}/\left(\lfloor M\rfloor-1\right)!=O\left(\log^{1+a}n\right) \tag{15}$$

and

$$\left(\frac{eu}{\lfloor M\rfloor-1}\right)^{\lfloor M\rfloor-1}=\exp\left(\left(\frac{1}{2}+a\right)\log\log_n+c_0+1+\{M\}+o\left(1\right)\right). \tag{16}$$

Now (12) implies that

$$\begin{aligned}S_{\lfloor M\rfloor-1}(u)=\frac{n}{u}\Bigg(1-\frac{1}{\sqrt{2\pi(\lfloor M\rfloor-1)}}\left(\frac{u}{\lfloor M\rfloor-1-u}\right)\\ \times\left(\frac{u}{\lfloor M\rfloor-1}\mathrm{e}^{1-u/(\lfloor M\rfloor-1)}\right)^{\lfloor M\rfloor-1}(1+O\left(1/M\right))\Bigg).\end{aligned} \tag{17}$$

Then, using (14)–(16) and some elementary calculations, we find that

$$f'\left(\xi\right)>cn \tag{18}$$

and

$$f\left(\rho\right)-f\left(u\right)=O\left(\log^{a+1}n\right) \tag{19}$$

for some $c>0$ and all sufficiently large n. Combining (11), (18), and (19), we obtain (10). This rough bound will be needed to prove the more precise estimates that follow. From the equations

$$n=\rho S_{\lfloor M\rfloor-1}\left(\rho\right),\qquad n=u\exp u,$$

we obtain

$$\log\left(\frac{u}{\rho}\right)=-u+\log\left(S_{\lfloor M\rfloor-1}\left(\rho\right)\right).$$

By Theorem 6, $S_{\lfloor M\rfloor-1}(\rho)=\mathrm{e}^{\rho}\cdot A$, where

$$\begin{aligned}A=1-\frac{1}{\sqrt{2\pi\left(\lfloor M\rfloor-1\right)}}\left(\frac{\rho}{\lfloor M\rfloor-1-\rho}\right)\\ \times\left(\frac{\rho}{\lfloor M\rfloor-1}\mathrm{e}^{1-\rho/(\lfloor M\rfloor-1)}\right)^{\lfloor M\rfloor-1}(1+O\left(1/M\right)).\end{aligned}$$

Hence

$$\rho-u=-\log A+\log\left(\frac{u}{\rho}\right). \tag{20}$$

By (10), however, we have

$$\log\left(\frac{u}{\rho}\right)=\log\left(1+\frac{u-\rho}{\rho}\right)=\frac{u-\rho}{\rho}+O\left(\log^{2a}n/n^2\right)=O\left(\log^{a}n/n\right). \tag{21}$$

Because $-\log(1-x) = x + O(x^2)$ as $x \to 0^+$, we have

$$\begin{aligned} -\log A &= \frac{1}{\sqrt{2\pi\left(\lfloor M \rfloor - 1\right)}} \left(\frac{\rho}{\lfloor M \rfloor - 1 - \rho} \right) \\ &\quad \times \left(\frac{\rho}{\lfloor M \rfloor - 1} \mathrm{e}^{1-\rho/(\lfloor M \rfloor - 1)} \right)^{\lfloor M \rfloor - 1} \left(1 + O\left(1/M\right)\right) \qquad (22) \\ &= \frac{\mathrm{e} \log^{a+1} n}{n} \exp\left(\{M\} + o\left(1\right)\right). \end{aligned}$$

Combining (20), (21) and (22), we obtain

$$\rho = u + \frac{\mathrm{e}\log^{a+1} n}{n} \exp\left(\{\mu_n + x\} + o\left(1\right)\right).$$

With Proposition 1 at our disposal, it is straightforward to make (8) explicit. First, from (9), (10), and (15) we see that $\Delta = \exp(o(1))$. Furthermore, we have

$$\frac{n}{\rho} - \frac{n}{u} = \frac{n\left(u - \rho\right)}{\rho u} = O\left(\log^{a-1} n\right).$$

Also, $n \log u - n \log \rho = n(u - \rho)/\rho + O(\log^{2a} n/n)$. By Proposition 1, this is

$$-\mathrm{e}\ \log^{a}\ n \exp\left(\{M\} + o\left(1\right)\right).$$

Finally, Stirling's formula, formula (9), and Proposition 1 imply that

$$\rho^{\lfloor M \rfloor} / \lfloor M \rfloor = \left(\mathrm{e} - 1\right)\left(\log^{a} n\right) \exp\left(\{M\} + o\left(1\right)\right).$$

Putting all these into (8), we obtain

$$\operatorname{Prob}\left(L_n \le \mu_n - a \log\log n\right) = \exp\left(-(\log^{a} n)\mathrm{e}^{\{\mu_n - a \log\log n\} + o(1)}\right)\left(1 + o\left(1\right)\right). \qquad \square$$

We omit the proof of Theorem 3 because it is essentially the same as that of Theorem 2. Sachkov's formula can also be proved this way. Our asymptotic estimates are not, however, uniform in x.

4. Nonexistence of a limit distribution for maximum block size. We claim that it is impossible to choose constants $\tilde{\mu}_n$ and $\tilde{\sigma}_n$ so that $(L_n - \tilde{\mu}_n)/\tilde{\sigma}_n$ has a nondegenerate limit distribution. Suppose that there were constants $\tilde{\mu}_n$ and $\tilde{\sigma}_n$ and a nondegenerate F such that, at any continuity point x,

$$\operatorname{Prob}\left(\frac{L_n - \tilde{\mu}_n}{\tilde{\sigma}_n} \le x\right) \to F(x) \quad \text{as } n \to \infty. \tag{23}$$

We must show that this leads to a contradiction.

Since F is nondegenerate, we can choose continuity points $\xi_1 < \xi_2$ such that $0 < F(\xi_i) < 1$. Then by (23)

$$\operatorname{Prob}\left(L_n \le \mu_n + \left(\tilde{\mu}_n - \mu_n + \xi_i \tilde{\sigma}_n\right)\right) \to F(\xi_i) \quad \text{as } n \to \infty. \tag{24}$$

Comparing (24) and Theorem 1, we see that $\tilde{\mu}_n - \mu_n + \xi_i \tilde{\sigma}_n$ is bounded ($i = 1, 2$). Hence $(\tilde{\mu}_n - \mu_n)$ and $\tilde{\sigma}_n$ are both bounded. We can therefore use (24) to conclude that

$$F(x) = \exp\left(-\exp\left(\mu_n - \tilde{\mu}_n - x\tilde{\sigma}_n + \{\tilde{\mu}_n + x\tilde{\sigma}_n\}\right)\right) + o(1). \tag{25}$$

Let $g_n(x) = \mu_n - \tilde{\mu}_n - x\tilde{\sigma}_n + \{\tilde{\mu}_n + x\tilde{\sigma}_n\}$. By (23), $g_n(x)$ converges. Hence $\{g_n(x)\}$ must either converge or accumulate at 0 and 1. To obtain the sought-after contradiction, we prove that, in fact, $\{g_n(x)\}$ is dense in $(0, 1)$. Toward this end, let $a_n = \tilde{\mu}_n + x\tilde{\sigma}_n$. Then $g_n(x) = \mu_n - (a_n - \{a_n\}) = \mu_n + k_n$ for some $k_n \in \mathbf{Z}$. Thus $\{g_n(x)\} = \{\mu_n\}$. It therefore suffices to prove the following result.

LEMMA 1. *The sequence of fractional parts $\langle\{\mu_n\}\rangle_{n=1}^{\infty}$ is dense in* $(0, 1)$.

Proof. Let $0 < a < b < 1$ be given. We must show that for some n_0, $\{\mu_{n_0}\} \in (a, b)$. Let $H(t) = \mathrm{e}U(t) - \log\sqrt{U(t)} - \log((\mathrm{e}-1)\sqrt{2\pi \mathrm{e}})$, where $U(t)$ is the positive solution to $z \exp z = t$. Both $H(t)$ and $U(t)$ are continuously differentiable, strictly increasing functions of t. In fact,

$$U'(t) = 1/\left(\mathrm{e}^{U(t)} + U(t)\,\mathrm{e}^{U(t)}\right) \quad \text{and} \quad H'(t) = \mathrm{e}U'(t) - \frac{U'(t)}{2U(t)}.$$

(To see that H is increasing for $t >$ some t_0, recall that $U \exp U = t$.) Now let K be the inverse of H; $K(H(t)) = t = H(K(t))$. It is not difficult to see that K exists and is differentiable. Furthermore,

$$K'(t) = 1/H'(K(t)) \to \infty \quad \text{as } t \to \infty.$$

It follows by the mean value theorem that

$$K(b + l) - K(a + l) \to \infty \quad \text{as } l \to \infty.$$

In particular, $K(b + l) - K(a + l) > 1$ for l sufficiently large, so there are positive integers l_0 and n_0 for which $K(a + l_0) < n_0 < K(b + l_0)$. Then, however,

$$H(K(a + l_0)) < H(n_0) < H(K(b + l_0)),$$

i.e., $a + l_0 < \mu_{n_0} < b + l_0$. Since l_0 is an integer, it follows that $\{\mu_{n_0}\} \in (a, b)$. This proves Lemma 1, thereby completing the proof of Theorem 5 as well. □

5. Multiplicity of the largest block size. This section is devoted to a proof of the following theorem.

THEOREM 7. *For any fixed positive integer m,*

$$\begin{aligned}&\operatorname{Prob}(R_n = m)\\ &= \frac{(2\pi)^{-m/2}}{m!} \exp\left(m\left(\{m_n\} - \frac{1}{2}\right)\right) \sum_{\xi=-\infty}^{\infty} \exp\left(\frac{-\mathrm{e}}{\sqrt{2\pi \mathrm{e}}\,(\mathrm{e}-1)}\,\mathrm{e}^{\{m_n\}-\xi} - m\xi\right) + o(1)\end{aligned}$$

as $n \to \infty$.

Note the presence of the fractional part $\{m_n\}$ in the limit of $\operatorname{Prob}(R_n = m)$. (For results closely related to Theorem 7, see Fristedt [8].)

Let $P(n, l, m)$ be the number of partitions of $[n]$ with exactly m blocks of size l and no blocks larger than l. There are $\binom{n}{lm}$ ways to choose the elements of the largest blocks and $(lm)!/l!^m m!$ ways to partition them into m blocks of size l. Hence, for $l > 1$,

$$P(n,l,m) = \binom{n}{lm} \frac{(lm)!}{l!^m m!} \Psi(n - lm, l - 1), \tag{26}$$

and consequently

$$\begin{aligned} \operatorname{Prob}(R_n = m) &= \frac{n!}{B(n)} \sum_{l=2}^{n/m} \frac{\Psi(n - lm, l - 1)}{l!^m m! (n - lm)!} + \frac{\delta_{m,n}}{B(n)} \\ &= \frac{n!}{eB(n)} \sum_{l=2}^{n/m} \frac{I_l}{l!^m m!} + \frac{\delta_{m,n}}{B(n)}, \end{aligned} \tag{27}$$

where

$$I_l = \frac{1}{2\pi i} \oint \frac{\exp(S_{l-1}(z))}{z^{n-lm+1}} dz$$

and $\delta_{m,n} = 1$ if $m = n$, zero otherwise.

This integral can certainly be estimated by the saddle point method. However, we cannot simply plug these estimates into (27); there appears to be a problem with uniformity. However, this problem can be circumvented via Proposition 2.

PROPOSITION 2. *For any $\varepsilon > 0$, there is a W_ε such that for all $m \in \mathbf{Z}^+$ and all $n > n_{\varepsilon,m}$,*

$$\left| \operatorname{Prob}(R_n = m) - \sum_{|l-m_n|<W_\varepsilon} \frac{P(n,l,m)}{B(n)} \right| < \varepsilon.$$

Proof. We have

$$\begin{aligned} \left| \operatorname{Prob}(R_n = m) - \sum_{|l-m_n|<W_\varepsilon} \frac{P(n,l,m)}{B(n)} \right| &= \sum_{|l-m_n|\geq W_\varepsilon} \frac{P(n,l,m)}{B(n)} \\ &\leq \sum_{|l-m_n|\geq W_\varepsilon} \operatorname{Prob}(L_n = l). \end{aligned}$$

By Theorem 1, we can choose W_ε sufficiently large so that for all large n, $\operatorname{Prob}(|L_n - m_n| \geq W_\varepsilon) < \varepsilon$. □

By Proposition 2, we need only estimate I_l for l in an interval around m_n of bounded (n-independent) width. Uniformity is therefore not an issue.

We emphasize that the procedure to estimate I_l is very similar to that of Theorem 2.

Let $g(z) = S_{l-1}(z) - (n - lm) \log z$ so that

$$I_l = \frac{1}{2\pi \mathrm{i}} \oint \exp(g(z)) \frac{dz}{z}.$$

Let $\alpha = \alpha(n,l,m)$ be the positive solution to

$$\alpha S_{l-2}(\alpha) = n - lm. \tag{28}$$

Then the saddle point method yields

$$I_l = \frac{\exp\left(S_{l-1}(\alpha)\right)}{\alpha^{n-lm+1}\sqrt{2\pi g''(\alpha)}}\left(1+o(1)\right). \tag{29}$$

To simplify the right side of (29), observe that

$$S_{l-1}(\alpha) = \frac{n-lm}{\alpha} + \frac{\alpha^{l-1}}{(l-1)!}.$$

Similarly,

$$g''(\alpha) = \frac{n-lm}{\alpha} - \frac{\alpha^{l-1}}{(l-2)!} + \frac{n-lm}{\alpha^2}.$$

Putting these back into (29), we obtain

$$I_l = \frac{\exp\left(\frac{n-lm}{\alpha} + \frac{\alpha^{l-1}}{(l-1)!}\right)}{\alpha^{n-lm+1}\sqrt{2\pi\frac{n-lm}{\alpha} - \frac{\alpha^{l-2}}{(l-2)!} + \frac{n-lm}{\alpha^2}}}\left(1+o(1)\right). \tag{30}$$

To make this more explicit, we must estimate α.

LEMMA 2. *Let $\xi = l - \lfloor m_n \rfloor$. Then, for $|\xi| \le W_\varepsilon$,*

$$\alpha = u + \frac{\mathrm{e}^2}{\sqrt{2\pi \mathrm{e}}\,(\mathrm{e}-1)}\exp\left(\{m_n\} - \xi\right)\frac{u}{n} - \frac{\mathrm{e}m\log n}{n} + o\left(\frac{\log n}{n}\right).$$

The proof of Lemma 2 is omitted because it is almost the same as the proof of Proposition 1. □

The following estimate is well known [5]:

$$\frac{n!}{B(n)} = \mathrm{e}\sqrt{2\pi}\exp\left(-\mathrm{e}^{u_n} + u_n\mathrm{e}^{u_n}\log u_n + \frac{1}{2}u_n + \log u_n\right)\left(1+o(1)\right). \tag{31}$$

Together with Lemma 2, this enables us to simplify each term in the exponent of (30). Thus

$$\frac{n!}{B(n)} I_l = \mathrm{e}\exp\Bigg(\left(\frac{n}{\alpha} - \frac{n}{u}\right) - \frac{lm}{\alpha} + \left(n\log u - n\log\alpha\right) + lm\log\alpha + \frac{\alpha^{l-1}}{(l-1)!}\Bigg)\left(1+o(1)\right),$$

where

$$\frac{n}{\alpha} - \frac{n}{u} = o(1),$$

$$-\frac{lm}{\alpha} = -m\mathrm{e} + o(1),$$

$$n\log u - n\log\alpha = \frac{-\mathrm{e}^2}{\sqrt{2\pi\mathrm{e}}\,(\mathrm{e}-1)}\mathrm{e}^{\{m_n\}-\xi} + \mathrm{e}m + o(1),$$

and

$$\frac{\alpha^{l-1}}{(l-1)!} = \frac{\mathrm{e}^{\{m_n\}-\xi+1}}{\sqrt{2\pi\mathrm{e}}} + o(1).$$

Hence

(32)
$$\frac{n!}{B(n)} I_l = \mathrm{e}\,\exp\left(\frac{-\mathrm{e}^2}{\sqrt{2\pi\mathrm{e}}\,(\mathrm{e}-1)}\,\mathrm{e}^{\{m_n\}-\xi} + \frac{1}{\sqrt{2\pi\mathrm{e}}}\,\mathrm{e}^{\{m_n\}-\xi+1} + lm\log\alpha + o(1)\right).$$

Observe that

$$\frac{1}{l!^m m!} = \frac{(2\pi)^{-m/2}}{m!}\exp\left(-\frac{m}{2} - lm\log u + m\{m_n\} - m\xi + o(1)\right). \tag{33}$$

Putting (32) and (33) into (26), we obtain

$$\sum_{|l-m_n|<W_\varepsilon} \frac{P(n,l,m)}{B(n)}$$
$$= \frac{(2\pi)^{-m/2}}{m!}\exp\left(m\left(\{m_n\}-\frac{1}{2}\right)\right)\sum_{|\xi|\le W_\varepsilon}\exp\left(\frac{-\mathrm{e}}{\sqrt{2\pi\mathrm{e}}\,(\mathrm{e}-1)}\,\mathrm{e}^{\{m_n\}-\xi} - m\xi\right) + o(1).$$

Note that the series

$$\sum_{\xi=-\infty}^{\infty}\exp\left(\frac{-\mathrm{e}}{\sqrt{2\pi\mathrm{e}}\,(\mathrm{e}-1)}\,\mathrm{e}^{\{m_n\}-\xi} - m\xi\right)$$

converges. Theorem 7 now follows from Proposition 2. □

It is probably possible to give a different proof of Theorem 7 using the results in [8]. Uniformity would again be the major issue.

6. The average multiplicity. The average multiplicity is

$$E(R_n) = \sum_{m=1}^{n} m\,\mathrm{Prob}(R_n = m). \tag{34}$$

We would like to use Theorem 7 to estimate m Prob$(R_n = m)$ and then sum. This, however, can only be justified if we truncate the sum after a finite (n-independent) number of terms. Therefore, uniformity is a major issue here. Thus, to prove Theorem 4, we must show that all but a bounded number of terms can be neglected. That is the purpose of Proposition 3.

PROPOSITION 3. *Let $\omega(n) \to \infty$ arbitrarily slowly. Then*

$$\sum_{m\ge\omega(n)} m\,\mathrm{Prob}(R_n = m) = o(1).$$

The proof of this key proposition is surprisingly difficult. We have

$$\sum_{m\ge\omega(n)} m\,\mathrm{Prob}(R_n = m) = \sum_{l=1}^{n}\sum_{m\ge\omega(n)}\left(m\,\mathrm{Prob}(R_n = m\,\mathrm{and}\,L_n = l)\right). \tag{35}$$

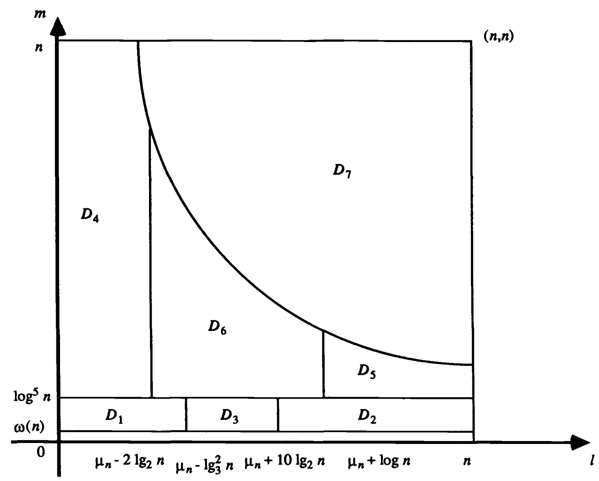


FIG. 1. *Seven domains of summation.*

The double sum on the right of (35) is broken up into seven domains as shown schematically in Fig. 1. Let $\lg_k n$ denote the k-times-iterated natural logarithm. Then

$$D_1 = \left\{(l,m)\colon\ \omega(n) \le m \le \log^5 n \,\text{and}\, 1 \le l \le \mu_n - \lg_3^2 n\right\},$$
$$D_2 = \left\{(l,m)\colon\ \omega(n) \le m \le \log^5 n \,\text{and}\, n \ge l \ge \mu_n + 10\lg_2 n\right\},$$
$$D_3 = \left\{(l,m)\colon\ \omega(n) \le m \le \log^5 n \,\text{and}\, \mu_n - \lg_3^2 n < l < \mu_n + 10\lg_2 n\right\},$$
$$D_4 = \left\{(l,m)\colon\ \log^5 n \le m \le n \,\text{and}\, 1 \le l < \mu_n - 2\lg_2 n \,\text{and}\, lm \le n - n/\lg_3 n\right\},$$
$$D_5 = \left\{(l,m)\colon\ \log^5 n \le m \le n \,\text{and}\, n \ge l > \mu_n + \log n \,\text{and}\, lm \le n - n/\lg_3 n\right\},$$
$$D_6 = \left\{(l,m)\colon\ \log^5 n \le m \le n \,\text{and}\, \mu_n - 2\lg_2 n \le l \le \mu_n + \log n \right.$$
$$\left. \text{and}\, lm \le n - n/\lg_3 n\right\},$$
$$D_7 = \left\{(l,m)\colon\ \log^5 n \le m \le n \,\text{and}\, \left(\frac{n - n/\lg_3 n}{m}\right) \le l \le n\right\}.$$

The proof of Proposition 3 therefore consists of seven lemmas: one for each domain. For $i = 1, 2, \ldots, 7$, let $S_i := \sum_{(l,m)\in D_i}(m \operatorname{Prob}(R_n = m \text{ and } L_n = l))$.

LEMMA 3. *It holds that* $S_1 = o(1)$.

Proof. We have

$$S_1 \le \sum_{l=1}^{\mu_n - \lg_3^2 n} \sum_{m\ge 1} (\log n)^5 \operatorname{Prob}(R_n = m \text{ and } L_n = l)$$
$$\le (\log n)^5 \operatorname{Prob}\left(L_n \le \mu_n - \lg_3^2 n\right).$$

Theorem 1 then implies that

$$S_1 = o(1)\,. \qquad \square \tag{36}$$

LEMMA 4. *It holds that* $S_2 = o(1)$.

Proof. Using Theorem 2 with $a = -10$, we have

$$S_2 \le (\log n)^5 \operatorname{Prob}(L_n \ge \mu_n + 10 \lg_2 n) = o(1)\,. \qquad \square \tag{37}$$

LEMMA 5. *It holds that* $S_3 = o(1)$.

Proof. Our estimates for S_3 begin with the observation that (see (26))

$$\begin{aligned} m\,\mathrm{Prob}\,(R_n = m \,\mathrm{and}\, L_n = l) &= m\,\frac{1}{B\,(n)}\binom{n}{lm}\frac{(lm)!B\,(n-lm)}{l!^m m!} \\ &\quad\times\left(\frac{\Psi\,(n-lm,l-1)}{B\,(n-lm)}\right) \\ &= \left\{\frac{n!}{B\,(n)}\right\}\left\{\frac{B\,(n-lm)}{(n-lm)!}\right\} \\ &\quad\times\left\{\frac{1}{l!^m\,(m-1)!}\right\}\mathrm{Prob}\,(L_{n-lm}\le l-1)\,. \end{aligned} \tag{38}$$

Then split S_3 into $S_3'+S_3''$, where $S_3' = \sum_{l=\mu_n-\mathrm{lg}_3^2 n}^{\mu_n}(*)$; and $S_3'' = \sum_{l=\lfloor\mu_n\rfloor+1}^{\mu_n+10\mathrm{lg}_2 n}(*)$. Our immediate goal is to obtain upper bounds for each of the four factors in (38). To yield a useful estimate for S_3', these bounds must be uniform for $\mu_n - \mathrm{lg}_3^2\, n \le l \le \mu_n$ and $\omega(n) \le m \le \log^5 n$. Let $N = N(n,l,m) = n - lm$. Recall the definition of u_N in the Introduction. As in §3, we use the mean value theorem to obtain

$$u - u_N = \frac{lm}{\xi e^\xi + e^\xi} \quad \text{for some } \xi \in [u_N, u]\,.$$

Thus

$$u - u_N \le \frac{lm}{n-lm} \quad \text{and} \quad u - u_N = O\left(\frac{\log^6 n}{n}\right) \tag{39}$$

uniformly for l and m in the range of summation S_3'. Combining this with Theorem 1, we can estimate the fourth factor of (38),

$$\mathrm{Prob}\,(L_{n-lm}\le l-1) = \exp\left(-\exp\left(-l+\mu_N+1\right)\right)\left(1+o\left(1\right)\right).$$

Using (39), we obtain

$$\mathrm{Prob}\,(L_{n-lm}\le l-1) = \exp\left(-\exp\left(-l+\mu_n+1\right)\right)\left(1+o\left(1\right)\right), \tag{40}$$

where the lowercase o constant holds uniformly.

By (31), there is an absolute constant $c > 0$ such that

$$\frac{B\,(n-lm)}{(n-lm)!} \le c\ \exp\left(-u_N \mathrm{e}^{u_N}\log u_N + \mathrm{e}^{u_N} - \frac{1}{2}\,u_N - \log u_N\right) \tag{41}$$

(for l and m in the range of S_3'). For the same reason, we have

$$\frac{n!}{B\,(n)} \le c\ \exp\left(u\mathrm{e}^u\log u - \mathrm{e}^u + \frac{1}{2}\,u + \log u\right). \tag{42}$$

Combining (41), (42), and Stirling's formula, we obtain

$$m\,\mathrm{Prob}\,(R_n = m \,\mathrm{and}\, L_n = l) \le \frac{c\exp\left(T_1+T_2\right)}{(2\pi)^{m/2}\,(m-1)!}\,\mathrm{Prob}\,(L_{n-lm}\le l-1)\,, \tag{43}$$

where

$$T_1 = lm \log u_N - lm \log\left(\frac{l}{\mathrm{e}}\right) - \frac{1}{2} m \log l$$

and

$$T_2 = n \log \frac{u}{u_N} + n\left(\frac{1}{u_N} - \frac{1}{u}\right) + \frac{u - u_N}{2} - \frac{lm}{u_N} + \log \frac{u}{u_N}.$$

Using (39), we can verify that the second term in the exponent of (43) is negligible, as follows:

$$T_2 = o(m). \tag{44}$$

To estimate T_1, let $x = l - \lfloor \mu_n \rfloor$ and observe that $-\mathrm{lg}_3^2\, n \le x \le 0$. By (39), we have

$$\begin{aligned} T_1 &= \left(lm \log u - lm \log\left(\frac{l}{e}\right)\right) - \frac{1}{2} m \log l + O\left(\frac{\log^{11} n}{n}\right) \\ &= lm \log\left(\frac{\mathrm{e}u}{l}\right) - \frac{m}{2} - \frac{m}{2} \log u + o(m). \end{aligned} \tag{45}$$

However,

$$\begin{aligned} lm \log\left(\frac{\mathrm{e}u}{l}\right) &= -lm\left(\log\left(1 - \frac{(\log u)/2 + c_0 - x + \{\mu_n\}}{eu}\right)\right) \\ &\le \frac{lm}{\mathrm{e}u}\left(\frac{1}{2}\log u + c_0 - x + \{\mu_n\}\right) + o(m) \\ &= \frac{m}{2}\log u + m\left(c_0 + \{\mu_n\}\right) - mx + o(m). \end{aligned} \tag{46}$$

Combining (43)–(46), we obtain

$$\begin{aligned} &m \,\mathrm{Prob}\,(R_n = m \mathrm{and} L_n = l) \\ &\qquad \le \frac{c \exp\left(m\left(c_0 - \frac{1}{2} + \{\mu_n\}\right) - m\left(l - \lfloor \mu_n \rfloor\right) + o(m)\right) \exp\left(-\mathrm{e}^{\mu_n - l + 1}\right)}{\left(\sqrt{2\pi}\right)^m (m-1)!} \end{aligned}$$

for all large n. Because this bound is uniform (for l and m in the range of S_3'), we can simply sum to estimate S_3'. Recall $c_0 = \log((\mathrm{e}-1)\sqrt{2\pi\mathrm{e}})$. Thus

$$\begin{aligned} S_3' &\le c \sum_{m=\omega(n)}^{\log^5 n} \sum_{l=\mu_n - \mathrm{lg}_3^2\, n}^{\mu_n} \frac{\exp\left(m(\log(\mathrm{e}-1) + \{\mu_n\}) - m(l - \lfloor \mu_n \rfloor) - \mathrm{e}^{\mu_n - l + 1} + o(m)\right)}{(m-1)!} \\ &= c \sum_{m=\omega(n)}^{\log^5 n} \left(\frac{\exp\left(m\left(\log(\mathrm{e}-1) + \{\mu_n\} + o(1)\right)\right)}{(m-1)!} \cdot \sum_{x=0}^{\mathrm{lg}_3^2\, n} \exp\left(mx - \mathrm{e}^{x+1+\{\mu_n\}}\right)\right). \end{aligned}$$

Now let $h(y) = my - \mathrm{e}^{y+1+\{\mu_n\}}$. Let x_0 be the point at which h attains its maximum, namely, $x_0 = \log m - 1 - \{\mu_n\}$. Then the inner sum above is

$$\sum_{x=0}^{\mathrm{lg}_3^2\, n} \mathrm{e}^{h(x)} = \sum_{x=0}^{x_0} \mathrm{e}^{h(x)} + \sum_{x=\lfloor x_0 \rfloor + 1}^{\mathrm{lg}_3^2\, n} \mathrm{e}^{h(x)}. \tag{47}$$

Crude estimates suffice to bound the first sum in (47), as follows:

$$\sum_{x=0}^{x_0} \mathrm{e}^{h(x)} \le (x_0+1)\,\mathrm{e}^{h(x_0)} \le c\,\log m \exp\left(m\log m - m\left(\{\mu_n\}+2\right)\right).$$

For the second sum, we use the fact that h is nonincreasing, shown below:

$$\sum_{x=\lfloor x_0\rfloor+1}^{\lg_3^2 n} \mathrm{e}^{h(x)} \le \int_{\lfloor x_0\rfloor}^{\infty} \mathrm{e}^{h(y)}\,dy < \mathrm{e}^{-m(1+\{\mu_n\})}\int_0^{\infty} y^{m-1}\mathrm{e}^{-y}\,dy \tag{48}$$
$$= \Gamma(m)\,\mathrm{e}^{-m(1+\{\mu_n\})}.$$

Using (47), (48), and Stirling's formula, we obtain

$$\sum_{x=0}^{\lg_3^2 n} \exp\left(mx - \mathrm{e}^{x+1+\{\mu_n\}}\right) = O\left(\log m \exp\left(m\log m - m\left(\{\mu_n\}+2\right)\right)\right).$$

Hence

$$S_3' = O\left(\sum_{m=\omega(n)}^{\log^5 n} \log m \exp\left(\left(\log\frac{\mathrm{e}-1}{\mathrm{e}} + o(1)\right)m\right)\right),$$

which is a tail of a convergent series. Thus $S_3' = o(1)$.

A similar argument works for S_3'', below:

$$\begin{aligned} S_3'' &\le \sum_{l=\mu_n}^{\mu_n+10\lg_2 n}\ \sum_{m=\omega(n)}^{\log^5 n} \left\{\frac{n!}{B(n)}\right\}\left\{\frac{B(n-lm)}{(n-lm)!}\right\}\left\{\frac{1}{l!^m\,(m-1)!}\right\} \\ &\le c\sum_{m=\omega(n)}^{\log^5 n}\left(\frac{\exp\left(m\left(c_0-\frac12+\{\mu_n\}+o(1)\right)\right)}{\left(\sqrt{2\pi}\right)^m(m-1)!}\cdot\sum_{l=\mu_n}^{\mu_n+10\lg_2 n}\exp\left(-m\left(l-\mu_n\right)\right)\right). \end{aligned}$$

Note that $\sum_{l=\mu_n}^{\mu_n+10\lg_2 n}\exp(-m(l-\mu_n))$ is uniformly bounded for all $m \ge 1$. Thus

$$S_3'' \le c\sum_{m=\omega(n)}^{\log^5 n}\frac{\exp\left(m\left(c_0-\frac12+\{\mu_n\}+o(1)\right)\right)}{\left(\sqrt{2\pi}\right)^m(m-1)!} = o(1).$$

This completes the proof of Lemma 5. □

LEMMA 6. *It holds that* $S_4 = o(1)$.

Proof. We have $S_4 \le \sum_{l<\mu_n-2\lg_2 n}\sum_{m\ge1}(*) \le n\,\mathrm{Prob}(L_n \le \mu_n - 2\lg_2 n)$. Taking $a=2$ in Theorem 2, we see that $S_4 = o(1)$. □

LEMMA 7. *It holds that* $S_5 = o(1)$.

Proof. Putting $b=1$ in Theorem 3, we obtain

$$S_5 \le n\,\mathrm{Prob}\left(L_n > \mu_n + \log n\right) = O\left(\frac{1}{n^{\cdot 15}}\right) = o(1). \qquad \square$$

LEMMA 8. *It holds that* $S_6 = o(1)$.

Proof. We have

$$S_6 \le \sum_{(l,m)\in D_6} \left\{\frac{n!}{B(n)}\right\}\left\{\frac{B(n-lm)}{(n-lm)!}\right\}\left\{\frac{1}{l!^m\,(m-1)!}\right\}.$$

To estimate this sum, we use (41), (42), and Stirling's formula. Thus

(49)

$$\begin{aligned}\frac{n!}{B(n)}\frac{B(n-lm)}{(n-lm)!}\frac{1}{l!^m(m-1)!} \le K\exp\Bigg(&n\log\frac{u}{u_N} + n\left(\frac{1}{u_N}-\frac{1}{u}\right)\\ &+\frac{u-u_N}{2} - \frac{lm}{u_N} + \log\frac{u}{u_N}\\ &+ \left(lm\log u_N - lm\log(l/\mathrm{e}) - (m\log l)/2\right)\Bigg)\\ &\cdot\frac{1}{\left(\sqrt{2\pi}\right)^m(m-1)!},\end{aligned}$$

where K is an absolute constant. To simplify (49), we note that

$$u - u_N \le \frac{lm}{n-lm} \le \frac{lm}{n-(n-n/\lg_3 n)} = \frac{lm\lg_3 n}{n}. \tag{50}$$

It is not too hard to see that

$$n\log\frac{u}{u_N} + n\left(\frac{1}{u_N}-\frac{1}{u}\right) + \frac{u-u_N}{2} - \frac{lm}{u_N}\log\frac{u}{u_N} \le Km\lg_3 n,$$

where K is absolute. Also,

$$lm\log u_N - lm\log(l/\mathrm{e}) - (m\log l)/2 \le 2m\lg_2 n.$$

Since summation with respect to l produces at most a factor of $\log n + 2\lg_2 n$, we have

$$S_6 \le K(\log n + 2\lg_2 n)\sum_{m\ge\log^5 n}\frac{\exp(2m\lg_2 n)}{\left(\sqrt{2\pi}\right)^m(m-1)!}.$$

Again, by Stirling's formula, we have

$$S_6 \le K(\log n + 2\lg_2 n)\sum_{m\ge\log^5 n}\exp(-m(5\lg_2 n - 2\lg_2 n)) = o(1). \qquad \square$$

LEMMA 9. *It holds that* $S_7 = o(1)$.

Proof. To estimate S_7 we note that the summation index l may be confined to the interval $\mu_n - 2\lg_2 n \le l \le \mu_n + \log n$. It therefore suffices to obtain an upper estimate for

$$\begin{aligned}S_7' :=& \sum_{\substack{(l,m)\in D_7\\ \mu_n-2\lg_2 n\le l\le\mu_n+\log n}} \left\{\frac{n!}{B(n)}\right\}\left\{\frac{B(n-lm)}{(n-lm)!}\right\}\left\{\frac{1}{l!^m\,(m-1)!}\right\}\\ \le& K\frac{n!}{B(n)}\sum_{\substack{(l,m)\in D_7\\ \mu_n-2\lg_2 n\le l\le\mu_n+\log n}}\frac{B(n-lm)}{(n-lm)!}\\ &\times\exp\left(-ml\log l + ml - \frac{m}{2}\log l - m\log m + m\right)\\ \le& K\frac{n!}{B(n)}\sum_{\substack{lm>n-n/\lg_3 n\\ m\ge(n-n/\lg_3 n)/(\mu_n+\log n)}}(*).\end{aligned} \tag{51}$$

It is straightforward to see that, for n sufficiently large, we have

$$-ml\log l + ml \le 2ml\,\frac{\lg_2 n}{\log n} \tag{52}$$

and

$$-m\log m + m \le -n/10. \tag{53}$$

Putting (52) and (53) in (51) yields

$$\begin{aligned} S_7' &\le K\frac{n!}{B(n)}\,n \sum_{n\ge y>n-n/\lg_3 n} \frac{B(n-y)}{(n-y)!}\exp\left(-y\log u + 2y\,\frac{\lg_2 n}{\log n} - n/10\right)\\ &= K\frac{n!}{B(n)}\,n \sum_{0\le t<n/\lg_3 n} \frac{B(t)}{t!}\exp\left(-(t-n)\log u + 2(n-t)\,\frac{\lg_2 n}{\log n} - n/10\right). \end{aligned}$$

Upon using (42) and the fact that

$$\frac{B(t)}{t!} \le K\exp\left(-t\log u_t + e^{u_t} - \frac{1}{2}\,u_t - \log u_t\right),$$

we obtain

$$S_7' \le Kn\,\exp\left(-\frac{n}{u} + \frac{u}{2} + \log u\right) \sum_{0\le t<n/\lg_3 n} \exp\left(t + 2n\,\frac{\lg_2 n}{\log n} - n/10\right).$$

Because of the term $-n/10$ in the exponent of the above sum, we clearly have $S_7' = o(1)$. Combining Lemmas 3–9, we obtain Proposition 3. □

7. Conclusions. The asymptotic distribution of the maximum block size depends on the fractional part of an implicitly defined root. This phenomenon is apparently quite common in asymptotic combinatorics and discrete probability theory. Most authors have simply assumed properties of the sequence of fractional parts, without studying them carefully.

Appendix. Let $\phi(t) = 1/t$ and let

$$r_\varepsilon(n) = \sqrt{2(1+\varepsilon)\log^+\left|\frac{\partial^2 h(n,\rho)}{\partial t^2}\right|}\Bigg/\left|\frac{\partial^2 h(n,\rho)}{\partial t^2}\right|^{1/2},$$

where $\log^+(x) := (\log x + |\log x|)/2$. To use Evgrafov's theorem, we must verify the following two conditions:

$$\frac{\phi(t)}{\phi(\rho)} \to 1 \quad \text{as}\, n \to \infty, \tag{A}$$

uniformly for $|t-\rho| < r_\varepsilon(n)$;

$$\frac{\partial^2 h(n,t)}{\partial t^2}\Bigg/\frac{\partial^2 h(n,\rho)}{\partial t^2} \to 1 \quad \text{as}\, n \to \infty, \tag{B}$$

uniformly for $|t-\rho| < r_\varepsilon(n)$. Using (6), we obtain $r_\varepsilon(n) = O(\log^{3/2} n/\sqrt{n})$. Then, however, for $|t-\rho| < r_\varepsilon(n)$, we have

$$\frac{\phi(t)}{\phi(\rho)} = \frac{1/t}{1/\rho} = \frac{\rho}{\rho + O\left(\log^{3/2} n/\sqrt{n}\right)} = 1 + o(1).$$

To verify (B), note that

$$\frac{\partial^2 h(n,t)}{\partial t^2} = S_{\lfloor M \rfloor -2}(t) + \frac{n}{t^2}.$$

Using Theorem 6 and Proposition 1, we obtain

$$\begin{aligned}\frac{\partial^2 h(n,t)}{\partial t^2} &= \mathrm{e}^t \left(1 - \frac{1}{\sqrt{2\pi(\lfloor M \rfloor - 2)}} \left(\frac{t}{\lfloor M \rfloor - 2 - t}\right)\right. \\ &\quad \times \left.\left(\frac{t}{\lfloor M \rfloor - 2} \mathrm{e}^{1-t/(\lfloor M \rfloor -2)}\right)^{\lfloor M \rfloor -2} (1 + O(1/M))\right) \\ &\quad + \frac{n}{t^2} = \mathrm{e}^t \left(1 + O\left(\log^{a+1} n/n\right)\right) + n/t^2\end{aligned}$$

for $|t-\rho| < r_\varepsilon(n)$. Hence

$$\frac{\partial^2 h(n,t)}{\partial t^2} \Bigg/ \frac{\partial^2 h(n,\rho)}{\partial t^2} = \frac{\mathrm{e}^t \left(1 + O\left(\log^{a+1} n/n\right)\right) + n/t^2}{\mathrm{e}^\rho \left(1 + O\left(\log^{a+1} n/n\right)\right) + n/\rho^2} = 1 + o(1),$$

uniformly for $|t-\rho| < r_\varepsilon(n)$.

Acknowledgment. We are grateful to Boris Pittel for calling our attention to Sachkov's paper [16].

REFERENCES

[1] R. Arratia and S. Tavaré, *Independent process approximations for random combinatorial structures*, Adv. Math., 1993.

[2] E. A. Bender, *Central and local limit theorems applied to asymptotic enumeration*, J. Combin. Theory Ser. A, 15 (1971), pp. 91–111.

[3] E. A. Bender, A. M. Odlyzko, and L. B. Richmond, *The asymptotic number of irreducible partitions*, European J. Combinatorics, 6 (1985), pp. 1–6.

[4] E. R. Canfield, *On a problem of Rota,* Adv. Math., 9 (1978), pp. 1–10.

[5] N. G. de Bruijn, *Asymptotic Methods in Analysis*, Dover, New York, 1981.

[6] J. de Laurentis and B. Pittel, *Counting subsets of the random partition and Brownian motion process*, Stochastic Process. Appl., 15 (1983), pp. 155–177.

[7] M. A. Evgrafov, *Asymptotic Estimates and Entire Functions,* Gordon and Breach, New York, 1961.

[8] B. Fristedt, *The Structure of Random Partitions of Large Sets,* Univ. of Minnesota Mathematics Report, May 1987.

[9] W. Goh and E. Schmutz, *Gap free set partitions,* Random Structures Algorithms, 3 (1992), pp. 9–18.

[10] J. Haigh, *Random equivalence relations,* J. Combin. Theory Ser. A, 16 (1972), pp. 287–295.

[11] L. H. Harper, *Stirling behavior is asymptotically normal,* Ann. Math. Statist., 38 (1967), pp. 410–414.

[12] ———, *A continuous analogue of Sperner's problem,* Pacific J. Math., 118 (1985), pp. 411–425.

[13] L. Kuipers and H. Niederreitter, *Uniform Distribution of Sequences,* Wiley–Interscience, New York, 1974, p. 53.

[14] L. Moser and M. Wyman, *An asymptotic formula for the Bell numbers,* Trans. Roy. Soc. Canada, 49 (1955), pp. 49–53.

[15] A. Odlyzko and L. B. Richmond, *On the number of distinct block sizes in partitions of a set,* J. Combin. Theory Ser. A, 38 (1985), pp. 170–181.

[16] V. N. Sachkov, *Random partitions of sets,* Theory Probab. Its Appl., 19 (1974), pp. 184–190.

[17] R. S. Varga, *Topics in Polynomial and Rational Interpolation and Approximation,* Les Presses de l'Université de Montréal, 1982.

[18] H. Wilf, *Generatingfunctionology,* Academic Press, New York, 1990.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 437–442, August 1994

STEINER DISTANCE-HEREDITARY GRAPHS*

D. P. DAY[†], ORTRUD R. OELLERMANN[‡], AND HENDA C. SWART[‡]

Abstract. Let G be a connected graph and $S \subseteq V(G)$. Then the Steiner distance of S in G, denoted by $d_G(S)$, is the smallest number of edges in a connected subgraph of G that contains S. A connected graph G is k-Steiner distance-hereditary, $k \geq 2$, if, for every $S \subseteq V(G)$ such that $|S| = k$ and every connected induced subgraph H of G containing $S, d_H(S) = d_G(S)$. It is shown that if G is 2-Steiner distance-hereditary, then G is k-Steiner distance-hereditary for all $k \geq 2$. Furthermore, it is shown that if G is k-Steiner distance-hereditary ($k \geq 3$), then G need not be $(k-1)$-Steiner distance-hereditary. An efficient algorithm for determining the Steiner distance of a set of k vertices in a k-Steiner distance-hereditary graph is discussed, and a characterization of 2-Steiner distance-hereditary graphs that leads to an efficient algorithm for testing whether a graph is 2-Steiner distance-hereditary is given.

Key words. Steiner distance, Steiner distance-hereditary graphs

AMS subject classification. 05C12

For graph theory terminology, we follow [1]. In particular, if G is a graph, then $V(G)$ and $E(G)$ denote the vertex and edge sets of G, respectively; we refer to $|V(G)|$ and $|E(G)|$ as the order and size of G, respectively. The *distance* $d_G(u,v)$ between two vertices u, v of a connected graph G is the length of a shortest $u-v$ path of G. The *eccentricity* $e(v)$ of a vertex v is $\max\{d(v,u)|u \in V(G)\}$. If G is a connected graph and $S \subseteq V(G)$, then the *Steiner distance* $d_G(S)$ is the size of a smallest connected subgraph of G that contains S. Such a subgraph is obviously a tree and is called a *Steiner tree* for S. If T is a tree, then a vertex of degree 1 in T is an *endvertex*, while all other vertices of T are called *internal* vertices of T.

Howorka [6] in 1977 defined a graph G to be *distance-hereditary* if each connected induced subgraph F of G has the property that $d_F(u,v) = d_G(u,v)$ for each $u, v \in V(F)$. To state the characterizations of distance-hereditary graphs given by Howorka [6], we need the following terminology. An *induced path* of G is a path that is an induced subgraph of G. Let $u, v \in V(G)$. Then a $u-v$ *geodesic* is a shortest $u-v$ path. Let C be a cycle of G. A path P is an *essential part* of C if P is a subgraph of C and $\frac{1}{2}|E(C)| < |E(P)| < |E(C)|$. An edge of G that joins two vertices of C that are not adjacent in C is called a *diagonal* of C. We say that two diagonals e_1, e_2 of C are *skew diagonals* if $C + e_1 + e_2$ is homeomorphic with K_4.

THEOREM A (Howorka). *The following are equivalent:* (i) *G is distance-hereditary;* (ii) *every induced path of G is a geodesic;* (iii) *no essential part of a cycle is induced;* (iv) *each cycle of length at least* 5 *has at least two diagonals, and each* 5*-cycle has a pair of skew diagonals;* (v) *each cycle of G of length at least* 5 *has a pair of skew diagonals.*

The definition of the Steiner distance of a set of vertices, together with the concept of distance-hereditary graphs, suggests a generalization of Steiner distance-hereditary graphs. A connected graph is *k-Steiner distance-hereditary,* $k \geq 2$ if, for every connected induced subgraph H of G of order at least k and set S of k ver-

*Received by the editors January 21, 1992; accepted for publication (in revised form) July 15, 1993.

[†] Technikon Natal, Durban, South Africa.

[‡] University of Natal, Durban, South Africa.

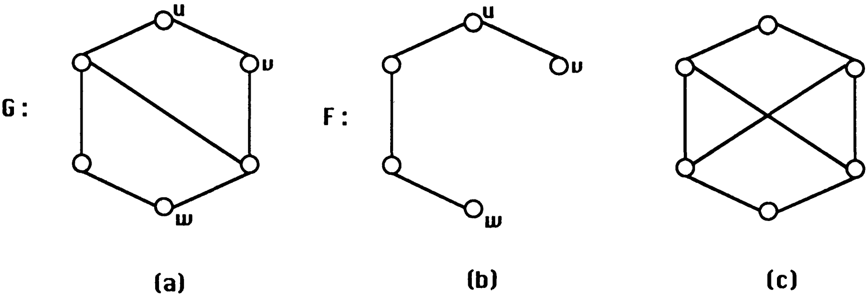


FIG. 1

tices of $H, d_H(S) = d_G(S)$. Thus 2-Steiner distance-hereditary graphs are distance-hereditary. Figure 1(a) shows a graph G that is not 3-Steiner distance-hereditary, since $d_F(\{u,v,w\}) \neq d_G(\{u,v,w\})$, where F is the induced subgraph of G shown in Fig. 1(b). However, it is not difficult to show that the graph of Fig. 1(c) is 3-Steiner distance-hereditary.

The problem of determining the Steiner distance of a set of vertices in a graph appears to be difficult. In fact, the following related decision problem π is NP-complete (see [4, p. 208]).

π: Suppose that G is a weighted graph whose edges have positive integer weights. Let $S \subseteq V(G)$ and suppose that B is a positive integer. Does there exist a subtree T of G that includes S and is such that the sum of the weights of the edges of T is no more than B?

Furthermore, the problem remains NP-complete even if G is a graph. This suggests solving the problem in certain special cases. If it is known that a graph is k-Steiner distance-hereditary, then $d_G(S)$ can easily be determined for every set S of $k \geq 2$ vertices of G as follows.

Let the vertices of $G - S$ be denoted by $v_1, v_2, \ldots, v_{p-k}$. Let $G_0 = G$. For each i $(1 \leq i \leq p-k)$, if the vertices of S belong to the same component of $G_{i-1} - v_i$, then G_i is defined to be $G_{i-1} - v_i$; otherwise, let G_i be G_{i-1}. Thus G_{p-k} is a connected induced subgraph of G that contains S. Therefore $d_{G_{p-k}}(S) = d_G(S)$. However, since the deletion of any vertex of G_{p-k} separates at least two vertices of S, no subgraph with fewer vertices than $p(G_{p-k})$ contains S and is connected. Thus G_{p-k} is a connected subgraph of smallest order that contains S. Hence any spanning tree of G_{p-k} is a Steiner tree for S.

Our first result shows that if G is a connected distance-hereditary graph, then $d_G(S)$ can be determined by the above procedure for any set $S \subseteq V(G)$ of at least two vertices.

THEOREM 1. *If G is 2-Steiner distance-hereditary, then G is k-Steiner distance-hereditary for $k \geq 3$.*

Proof. Suppose, to the contrary, that there exists a graph G that is 2-Steiner distance-hereditary, but not k-Steiner distance-hereditary for some $k \geq 3$. Let k be as small as possible and let H be a connected induced subgraph of G of smallest order, n say, for which there is a set S of k vertices of H such that $d_H(S) > d_G(S)$. Let $S = \{x_1, x_2, \ldots, x_k\}$. If $|V(H)| = k$, then there exists exactly one set of k vertices in H, namely $V(H)$. However, then every spanning tree of H is a Steiner tree for $V(H)$ in H and has size $k-1$. Since $d_G(V(H)) \geq k-1$, it follows that $d_G(V(H)) = d_H(V(H))$ in this case. This contradicts our choice of H. Hence $|V(H)| \geq k+1$. If $d_H(S) \leq k-2$, let T be a Steiner tree for S in H and let $H' = \langle V(T) \rangle_G$. Then $d_{H'}(S) = d_H(S) > d_G(S)$ and $|V(H')| < |V(H)|$, which contradicts our choice of H. Hence $d_H(S) = n-1$;

i.e., a Steiner tree for S in H must contain all the vertices of H. By our choice of $k, d_H(S-\{x_i\}) = d_G(S-\{x_i\})$ for all i $(1 \leq i \leq k)$.

We now show that no Steiner tree T' for S in G contains any x_i $(1 \leq i \leq k)$ as an internal vertex. Suppose that T' contains some x_i as internal vertex. Let $T_1, T_2, \ldots, T_m$ be the components of $T' - x_i$. Let T_1' be the subgraph of T' induced by $V(T_1) \cup \{x_i\}$ and let T_2' be the subgraph of T' induced by $(\bigcup_{j=2}^{m} V(T_j)) \cup \{x_i\}$. Let $S_1 = S \cap V(T_1')$ and $S_2 = S \cap V(T_2')$. Since $2 \leq |S_i| < k$ for $i = 1, 2$, it follows that $d_H(S_i) = d_G(S_i)$ for $i = 1, 2$. Furthermore, $|E(T_i')| = d_G(S_i)$ for $i = 1, 2$; otherwise, we can find a tree with fewer than $q(T') = d_G(S)$ edges that contains S. This is not possible. Let T_i be a Steiner tree for S_i in H $(i = 1, 2)$. Then $d_H(S) \leq d_H(S_1) + d_H(S_2) = |E(T_1')| + |E(T_2')| = d_G(S)$. This again produces a contradiction to the choice of S. Hence every Steiner tree for S in G has k endvertices that are precisely the vertices of S. Thus $d_G(S - \{x_i\}) < d_G(S)$ for all i $(1 \leq i \leq k)$.

We prove next that every vertex of S has degree 1 in H and is therefore an endvertex of every Steiner tree for S in H.

Let $x_i \in S$ and note that every Steiner tree for $S - \{x_i\}$ in H does not contain x_i; otherwise, $d_H(S) = d_H(S - \{x_i\}) = d_G(S - \{x_i\}) < d_G(S)$, which contradicts the fact that $d_H(S) > d_G(S)$. Let T_i be a Steiner tree for $S - \{x_i\}$ in H. Denote by P_i a shortest path in H from x_i to $V(T_i)$ and note that every vertex in H occurs in $V(T_i) \cup V(P_i)$ for $1 \leq i \leq k$, since $d_H(S) = |V(H)| - 1$. So P_i contains at least one edge. If P_i contains an internal vertex, w say, and $\deg_H x_i \geq 2$, then x_i has a neighbour y in H that is contained in $V(T_i)$ and $y \notin V(P_i)$, which produces a contradiction as x_i, y is a path from x_i to $V(T_i)$, which is shorter than P_i. Hence, if $\deg_H x_i \geq 2$, then P_i has length 1. Therefore

$$\begin{aligned} d_H(S) &= d_H(S - \{x_i\}) + 1 \\ &= d_G(S - \{x_i\}) + 1 \\ &\leq d_G(S), \end{aligned}$$

contrary to our assumption. Hence every $x_i \in S$ has degree 1 in H. Therefore every Steiner tree for S in H has k endvertices.

Next, consider T, a Steiner tree for S in H. Let l_i be the length of a shortest path Q_i (in H) from x_i to a vertex v_i of degree at least 3 in T for $i = 1, 2, \ldots, k$. Let $w_{i,1}$ be the vertex that precedes v_i on Q_i and observe that, with the possible exception of $w_{i,1}$, no internal vertex of Q_i has degree exceeding 2 in H. We now show that

$$(1) \qquad d_H(S) = \begin{cases} d_H(S - \{x_i\}) + l_i & \text{if } v_i \in V(T_i), \\ d_H(S - \{x_i\}) + l_i + 1 & \text{if } v_i \notin V(T_i), \end{cases}$$

where T_i is a Steiner tree on $S - \{x_i\}$ and where, in the latter case, $w_{i,1}$ has degree 2 in H.

We show first that $d_H(S - \{x_i\}) \geq d_H(S) - (l_i + 1)$. If this is not the case, then $d_H(S - \{x_i\}) \leq d_H(S) - l_i - 2$, and neither v_i nor any of its neighbours in T belongs to T_i. Let $w_{i,2}$ and $w_{i,3}$ be two vertices distinct from $w_{i,1}$ that are adjacent with v_i in T. Then $T - v_i w_{i,2}$ must contain x_i and $w_{i,3}$ in the same component, and thus some vertex $x_j \neq x_i$ such that the $x_i - x_j$ path P' in T contains $w_{i,3}$. Then P', together with T_i, produces a connected subgraph of H that contains S but not $w_{i,2}$. However, then $d_H(S) < p(H) - 1$, a contradiction. Hence $d_H(S - \{x_i\}) \geq d_H(S) - (l_i + 1)$.

If $v_i \in V(T_i)$, then the length of a shortest path from x_i to T_i is at most l_i. On the other hand, we know that it is at least l_i. Hence it is exactly l_i. So $d_H(S) = d_H(S - \{x_i\}) + l_i$ in this case. If $v_i \notin V(T_i)$, then some neighbour of v_i distinct from

$w_{i,1}$ must belong to T_i. Furthermore, v_i must be on a shortest path from x_i to T_i. Therefore $w_{i,1}$ has degree 2 in H. Hence $d_H(S) = d_H(S - \{x_i\}) + l_i + 1$ in this case.

Let T' be a Steiner tree for S in G and let $H' = \langle V(T')\rangle_G$. Since T' has k endvertices, there is some pair x_i, x_j of vertices of S for which the $x_i - x_j$ path in T' contains exactly one vertex of degree at least 3 in T', say y. Without loss of generality, we may assume that $x_i = x_1$ and $x_j = x_2$. Let $l_1' = d_{T'}(x_1, y)$ and $l_2' = d_{T'}(x_2, y)$. Observe that $d_G(x_1, x_2) \leq l_1' + l_2'$. Observe that $d_H(x_1, x_2) \geq l_1 + l_2 - 1$. Hence $d_G(x_1, x_2) \geq l_1 + l_2 - 1$. We now consider two cases.

Case 1. Suppose that $d_H(x_1, x_2) = l_1 + l_2 - 1$. Then $w_{1,1}$ and $w_{2,1}$ must be adjacent in H, and, furthermore, v_i must belong to T_i for $i = 1, 2$, by (1). Thus $d_H(S) = d_H(S - \{x_i\}) + l_i > d_G(S) \geq d_G(S - \{x_i\}) + l_i'$ for $i = 1, 2$. Therefore $l_i \geq l_i' + 1$ for $i = 1, 2$. Hence

$$d_H(x_1, x_2) = l_1 + l_2 - 1 \geq l_1' + l_2' + 1 > d_G(x_1, x_2),$$

a contradiction, since G is 2-Steiner distance-hereditary and because H is a connected induced subgraph of G.

Case 2. Suppose that $d_H(x_1, x_2) \geq l_1 + l_2$. Suppose first that $d_H(x_1, x_2) \geq l_1 + l_2 + 1$. Since $d_H(S - \{x_i\}) + l_i + 1 \geq d_H(S) > d_G(S) \geq d_G(S - \{x_i\}) + l_i'$, it follows that $l_i \geq l_i'$ for $i = 1, 2$. Hence $d_H(x_1, x_2) \geq l_1 + l_2 + 1 > l_1' + l_2' \geq d_G(x_1, x_2)$. This again contradicts the fact that G is 2-Steiner distance-hereditary. Suppose thus that $d_H(x_1, x_2) = l_1 + l_2$. Then $w_{1,1}$ and $w_{2,1}$ are not adjacent in H. If $d_H(S - \{x_i\}) + l_i = d_H(S)$ for $i = 1, 2$, then, by (1), v_i is in the vertex set of T_i. Suppose that $d_H(S - \{x_1\}) + l_1 = d_H(S)$. Then, as before, $l_1 \geq l_1' + 1$, and $l_2 \geq l_2'$. Hence $d_H(x_1, x_2) = l_1 + l_2 > l_1' + l_2' \geq d_G(x_1, x_2)$. This is not possible, since G is 2-Steiner distance-hereditary. So we may assume that $d_H(S) = d_H(S - \{x_i\}) + l_i + 1$ for $i = 1, 2$. Thus, by (1), $v_i \notin V(T_i)$ for $i = 1, 2$. We show next that $w_{1,1}$ and $w_{2,1}$ both have degree 2 in H. Suppose that $w_{1,1}$ has degree at least 3 in H. Let w be a vertex adjacent with $w_{1,1}$ that does not belong to Q_1. Then there is a path P in H from x_1 to T_1 that passes through w but does not contain v_1. Thus T_1, together with P, produces a connected subgraph of H that contains all the vertices S but not v_1. Thus $d_H(S) < p(H) - 1$, a contradiction. Therefore $w_{1,1}$ and $w_{2,1}$ both have degree 2 in H. Thus $v_1 = v_2$. However, then necessarily v_1 $(=v_2)$ must belong to T_1, so that $d_H(S) = d_H(S - \{x_1\}) + l_1$, which we have already shown cannot happen. □

Observe that for $k \geq 3$, the $(k+2)$-*cycle* C_{k+2} is $(k+2)$-, $(k+1)$-, and k-Steiner distance-hereditary, but not $(k-1)$-Steiner distance-hereditary. Thus the converse of Theorem 1 does not hold.

Several characterizations of distance-hereditary graphs that yield polynomial algorithms that test whether a graph is distance hereditary have been established. To state some of these characterizations, we define an *isolated vertex* to be a vertex having degree 0, and two vertices v and v' are *twins* if they have the same neighbourhood or the same closed neighbourhood.

The following characterization of distance-hereditary graphs was discovered independently by Bandelt and Mulder [1], D'Atri and Moscarini [3], and Hammer and Maffray [5].

THEOREM B. *A graph G is distance-hereditary if and only if every induced subgraph of G contains an isolated vertex, an endvertex, or a pair of twins.*

The result we establish next is another characterization of 2-Steiner distance-hereditary graphs and also suggests an efficient algorithm for determining whether a connected graph is 2-Steiner distance-hereditary. This result is also a direct consequence of a characterization of distance-hereditary graphs obtained independently by

Bandelt and Mulder [1] and D'Atri and Moscarini [3]. We need the following terminology. Suppose that G is a connected graph and that $u \in V(G)$. Let $V_{u,i} = \{x \in V(G) | d_G(u,x) = i\}$ for $0 \leq i \leq e_G(u)$, where $e_G(u)$ is the eccentricity of u in G, and let $N_{i-1}(u,v) = N(v) \cap V_{u,i-1}$ for $1 \leq i \leq e_G(u)$.

THEOREM 2. *A connected graph G contains an induced path that is not a geodesic if and only if there exists a vertex u and an integer $i \geq 2$ such that, for some pair x, y of vertices in $V_{u,i}$,*

(1) *$xy \in E(G)$ and $N_{i-1}(u,x) \neq N_{i-1}(u,y)$; or*

(2) *$xy \notin E(G), N_{i-1}(u,v) \neq N_{i-1}(u,y)$, and x and y are both adjacent with some vertex z in $V_{u,i+1}$.*

Proof. Suppose that there is some vertex u and an integer $i \geq 2$ such that, for some pair $x, y \in V_{u,i}$, (1) or (2) holds. Suppose first that (1) holds. Since $N_{i-1}(u,x) \neq N_{i-1}(u,y)$, so $N_{i-1}(u,x) - N_{i-1}(u,y) \neq \emptyset$ or $N_{i-1}(u,y) - N_{i-1}(u,x) \neq \emptyset$. Suppose that the former holds. Let $x_1 \in N_{i-1}(u,x) - N_{i-1}(u,y)$. Let P_1 be a shortest $u - x$ path that passes through x_1 and let P_2 be a shortest $u - y$ path. Let a be the last vertex that P_1 and P_2 have in common (possibly $a = u$). Then the vertices on the $a - x$ subpath of P_1, together with y, induce an $a - y$ path P that is longer than the $a - y$ subpath of P_2. Hence G contains an induced path that is not a geodesic.

Suppose now that (2) holds. We may again assume that there exists a vertex $x_1 \in N_{i-1}(u,x) - N_{i-1}(u,y)$. Clearly, $x_1 y \notin E(G)$ and $x_1 z \notin E(G)$. As above, let P_1 be a shortest $u-x$ path that contains x_1, P_2 a shortest $u-y$ path, and a the last vertex that P_1 and P_2 have in common. The vertices on the $a - x$ subpath of P_1, together with z and y, induce a path that has length two bigger than the $a - y$ subpath of P_2 (which is a geodesic). Hence G contains an induced subpath that is not a geodesic.

Conversely, suppose that G contains an induced path P (say a $u-v$ path) that is not a geodesic. Then $d_G(u,v) > 1$. Among the induced paths that are not geodesics, let P be as short as possible. We show that P has length at most $d_G(u,v)+2$. Suppose that $|E(P)| > d_G(u,v)+2$. Let $P : u = u_1, u_2, \ldots, u_n = v$. Then $d_G(u, u_{n-1}) \leq d_G(u,v)+1$, and $P' : u_1, u_2, \ldots, u_{n-1}$ is a path of length at least $|E(P)| - 1 \geq d_G(u,v) + 2 > d_G(u, u_{n-1})$. However, then P' is an induced path that is not a geodesic but has length less than P. This contradicts our choice of P. Hence P has length $d_G(u,v) + 1$ or $d_G(u,v) + 2$. Note that the $u - u_{n-1}$ subpath of P must be a geodesic; otherwise, we have a contradiction to our choice of P.

Thus, if $|E(P)| = d_G(u,v) + 1$, then $d_G(u, u_{n-1}) = d_G(u,v)$. Since the vertex that precedes u_{n-1} on P is not adjacent with $v, N_{i-1}(u, u_{n-1}) \neq N_{i-1}(u,v)$, where $i = d_G(u,v)$. If we let $x = u_{n-1}$ and $y = v$, then it follows that (1) holds.

Suppose now that $|E(P)| = d_G(u,v) + 2$. Then $d_G(u, u_{n-1}) = d_G(u, u_{n-1}) + 1$. Let x be u_{n-2} and let $y = v$. Then $xy \notin E(G)$, and the vertex that precedes x on P is not adjacent with y. Thus, if $i = d_G(u,v)$, then $N_{i-1}(u,x) \neq N_{i-1}(u,y)$. If we now let $z = u_{n-1}$, then $z \in V_{u,i+1}$, and z is adjacent with both x and y. Thus (2) holds. □

This result suggests a polynomial algorithm, using a breadth-first search technique that has complexity $O(|V(G)|^4)$ for determining whether a (connected) graph is 2-Steiner distance-hereditary. Spinrad [7] has developed an algorithm based on this characterization, which has complexity $O(|V(G)|^2)$. Once this is done and the graph has been found to be 2-Steiner distance-hereditary, we can efficiently determine, by Theorem 1, the Steiner distance of any set of vertices, which was also shown independently in [3].

In closing, we conjecture that, whenever G is k-Steiner distance-hereditary, then G is $(k + 1)$-Steiner distance-hereditary for $k \geq 3$.

REFERENCES

[1] H. J. Bandelt and H. M. Mulder, *Distance-hereditary graphs*, J. Combin. Theory Ser. B, 41 (1986), pp. 183–208.

[2] G. Chartrand and L. Lesniak, *Graphs and Digraphs*, Wadsworth & Brooks/Cole, Monterey, CA, 1986.

[3] A. D'Atri and M. Moscarini, *Distance-hereditary graphs, Steiner trees and connected domination*. SIAM J. Comput., 17 (1988), pp. 521–538.

[4] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, New York, 1979.

[5] P. L. Hammer and F. Maffray, *Completely separable graphs*, Discrete Appl. Math., 27 (1990), pp. 85–100.

[6] E. Howorka, *A characterization of distance hereditary graphs*, Quart. J. Math. Oxford, 28 (1977), pp. 417–420.

[7] J. Spinrad, *Prime testing for split decomposition of a graph*, SIAM J. Discrete Math., 2 (1989), pp. 590–599.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 443–457, August 1994

009

AN UPPER BOUND ON THE DIAMETER OF A GRAPH FROM EIGENVALUES ASSOCIATED WITH ITS LAPLACIAN*

F. R. K. CHUNG†, V. FABER‡, AND THOMAS A. MANTEUFFEL‡

Abstract. The authors give a new upper bound for the diameter $D(G)$ of a graph G in terms of the eigenvalues of the Laplacian of G. The bound is

$$D(G) \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right)} \right\rfloor + 1,$$

where $0 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of the Laplacian of G and where $\lfloor\ \rfloor$ is the floor function.

Key words. Laplacian, diameter, eigenvalues

AMS subject classification. 05C

1. Introduction.

Suppose that G is a connected graph (undirected or directed). For two vertices u and v in G, the distance between u and v, denoted by $d(u, v)$, is the length of a shortest path joining u and v in G. The diameter of G, denoted by $D(G)$, is the maximum distance over all pairs of vertices in G. The diameter is one of the graph variants that are not only of theoretical interest but also have many practical applications. In many communication models, diameter plays a key role in performance and cost optimization in network design when the time delay or signal degradation is proportional to the number of links that a message must travel. Numerous applications include circuit design, data representation, and parallel and distributive computing.

Let A denote the adjacency matrix of G. That is, $(A)_{ij} = 1$ if there is an edge between i and j (or from i to j), and it is 0 otherwise. Let d_i denote the degree of the ith vertex. When all d_i's are equal to k, we say that G is k-regular. The Laplacian of G is defined to be the matrix $Q = Q(G)$, where $(Q)_{ii} = d_i$ and $(Q)_{ij} = -A_{ij}$ if $i \neq j$. When a directed graph has the property that the in-degree is equal to the out-degree at every vertex, we can define the Laplacian analogously. The smallest nonzero eigenvalue, denoted by $\lambda(G)$, of $Q(G)$ can be used to derive various properties of G. It was shown in [12], [13] that $\lambda(G)$ provides useful bounds for the expanding properties of G. Using the expanding properties, Alon and Milman [1] deduced the following upper bound for the diameter for graphs with maximum degree k: $D(G) \leq 2\sqrt{2k/\lambda}\log_2 n$.

This bound was later improved [3] to $D(G) \leq \lceil \log(n-1)/\log(k/\bar{\lambda}) \rceil$, where $\bar{\lambda}$ denotes the second largest eigenvalue (in absolute value) of A. In this paper, we show that $D(G) \leq \lfloor \cosh^{-1}(n-1)/\cosh^{-1}(k/\bar{\lambda}) \rfloor + 1$, which is a special case of the following bound for general graphs:

$$D(G) \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right)} \right\rfloor + 1,$$

[*] Received by the editors June 19, 1991; accepted for publication (in revised form) March 23, 1993.

[†] Bell Communications Research, Morristown, New Jersey 07960.

[‡] Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

where $0 \leq \lambda_2 \leq \cdots \leq \lambda_n$ denote the eigenvalues of the Laplacian of G.

Some diameter bounds for directed graphs are also derived. Although these bounds are not as good as those for undirected cases, the techniques can be useful for bounding the diameter for Cayley graphs or on the factorization of groups. Namely, for a group H and a set S of generators, it is often desirable to find the least integer $m = D_S(H)$ such that every element of H can be written as the product of no more than m elements in S (repetitions are allowed). The upper bounds for diameters can then be applied for bounding $D_S(H)$ when the graph is considered to be the Cayley graph on H determined by S.

2. Diameter of a graph. Let G be a directed graph and A its adjacency matrix. Since the i, j entry of A^k is the number of directed paths from vertex i to vertex j of length k, the diameter D is the smallest natural number such that, given any i and j, one of the matrices in the set $\{A^m | 0 \leq m \leq D\}$ has a nonzero i, j entry. If a new matrix $\bar{A} = A - \Lambda$ is formed from A by subtracting any diagonal matrix Λ, the powers of $\bar{A}$ still have the same significance as the powers of A. To see this, note that $\bar{A}^m$ has a nonzero i, j entry only if there is a directed path of length at most m in G from vertex i to vertex j; if the distance from vertex i to vertex j in G is m, then $\bar{A}^m$ has a nonzero i, j entry. Thus, the diameter D is the smallest natural number such that, given any $i \neq j$, one of the matrices in the set $\{\bar{A}^m | 0 \leq m \leq D\}$ has a nonzero i, j entry.

We can now introduce polynomials into the definition of diameter. The statement that the i, j entry of at least one of the matrices in the set $\{\bar{A}^m | 0 \leq m \leq D\}$ is nonzero is equivalent to the statement that there exists a polynomial p_m of degree m less than or equal to D such that the i, j entry of $p_m(\bar{A})$ is nonzero. Thus the diameter D is the smallest natural number such that, given any $i \neq j$, there exists a polynomial p_m of degree $m \leq D$ such that the i, j entry of $p_m(\bar{A})$ is nonzero.

In fact, it can be easily seen that D is equal to the smallest natural number m for which there exists a polynomial p_m of degree m such that *all* off-diagonal entries of $p_m(\bar{A})$ are nonzero.

3. A fundamental inequality involving the Laplacian. In this section, it is essential to assume that the in-degree of each vertex of G is equal to its out-degree. We denote by d_i this common value for the vertex i. The Laplacian of G is defined to be $Q = \operatorname{diag}(d_i) - A$. Note that $Qu_1 = Q^*u_1 = 0$, where $u_1 = (1/\sqrt{n})(1, 1, \ldots, 1)^*$; so $QJ = JQ = 0$, where $J = nu_1u_1^*$ is the matrix all of whose entries are 1. Let e_i be the unit vector in the ith direction.

LEMMA 3.1. *Let B be an $n \times n$ matrix with the properties $Bu_1 = B^*u_1 = 0$. Then*

$$|B_{r,s}| \leq \|B\| \left(1 - \frac{1}{n}\right).$$

Proof. Define $f_i = e_i - (1/\sqrt{n})u_1$. Then $(f_r, u_1) = 0, (u_1, Bf_r) = 0$, and $\|f_r\| = \sqrt{1 - 1/n}$. Thus

$$|B_{r,s}| = |(e_r, Be_s)| = |(f_r, Bf_s)| \leq \|f_r\| \; \|B\| \; \|f_s\| = \|B\| \left(1 - \frac{1}{n}\right).$$

This proves the lemma. □

THEOREM 3.2. *Let $p_m(x)$ be a polynomial of degree m with $p_m(0) = 1$ such that*

$$\left\| p_m(Q) - \frac{J}{n} \right\| < \frac{1}{n - 1}.$$

Then $D \leq m$.

Proof. Let $B = p_m(Q) - J/n$. We have

$$Bu_1 = p_m(Q)\,u_1 - \frac{J}{n}u_1 = p_m(0)\,Iu_1 - u_1 = 0$$

and similarly $B^*u_1 = 0$. Thus Lemma 3.1 applies, and

$$\begin{aligned}(p_m(Q))_{r,s} &\geq \left(\frac{J}{n}\right)_{r,s} - |B_{r,s}| \\ &\geq \frac{1}{n} - \|B\|\left(1-\frac{1}{n}\right) \\ &> \frac{1}{n} - \frac{1}{n-1}\left(1-\frac{1}{n}\right) \\ &= 0.\end{aligned}$$

Thus, by our earlier remarks on diameter, $D \leq m$. This proves the theorem. □

4. The Chebychev polynomials.

The Chebychev polynomials of the first kind are given by

$$\begin{aligned}T_0(z) &= 1, \\ T_1(z) &= z, \\ T_{n+1}(z) &= 2zT_n(z) - T_{n-1}(z), \qquad n \geq 1.\end{aligned}$$

They may also be written as

$$T_n(z) = \cosh\left(n\cosh^{-1}(z)\right).$$

The Chebychev polynomials have the following interesting optimality property [14]. Let S_n be the set of all polynomials, $s_n(x)$, of degree n such that $s_n(0) = 1$. Let $[a,b]$ be an interval on the real line to the right of the origin. Then there exists a unique $t_n \in S_n$ such that

$$\max_{x\in[a,b]} |t_n(x)| = \min_{s_n\in S_n} \max_{x\in[a,b]} |s_n(x)|$$

and

$$t_n(x) = \frac{T_n\left(\frac{a+b-2x}{b-a}\right)}{T_n\left(\frac{a+b}{b-a}\right)}.$$

Furthermore,

$$\max_{x\in[a,b]} |t_n(x)| = t_n(a) = \frac{1}{T_n\left(\frac{a+b}{b-a}\right)}.$$

This result can be extended to a class of closed ellipses in the complex plane not containing the origin with foci lying on the real line or foci that are a complex conjugate pair. In particular, for $\delta > 0$ and $\gamma > 0$ real, if the ellipse E has center δ, foci $\delta \pm \gamma$, and semi-major axis of length α with $\gamma \leq \alpha < \delta$, then

$$t_n(z) = \frac{T_n\left(\frac{\delta - z}{\gamma}\right)}{T_n\left(\frac{\delta}{\gamma}\right)}$$

has the property

$$\max_{z\in E} |t_n(z)| = \frac{T_n\left(\frac{\alpha}{\gamma}\right)}{T_n\left(\frac{\delta}{\gamma}\right)}.$$

Furthermore, if $\gamma \leq \alpha \leq \alpha_n^* < \delta$, then

$$\max_{z\in E} |t_n(z)| = \min_{s_n\in S_n} \max_{z\in E} |s_n(z)|.$$

Here α_n^* is close to δ and depends on n in a complicated way. However, $\lim_{n\to\infty} \alpha_n^* = \delta$, and thus the Chebychev polynomials are asymptotically optimal [10], [6], [7]. If, on the other hand, the ellipse E has center δ, foci $\delta \pm i\gamma$, and semi-major axis of length $\alpha, \gamma \leq \alpha < (\delta^2+\gamma^2)^{1/2}$, then

$$t_n(z) = \frac{T_n\left(\frac{\delta-z}{i\gamma}\right)}{T_n\left(\frac{\delta}{i\gamma}\right)}$$

has real coefficients and for n even satisfies

$$\frac{T_n\left(\frac{\alpha}{\gamma}\right)}{T_n\left[\left(1+\left(\frac{\delta}{\gamma}\right)^2\right)^{1/2}\right]} = \max_{z\in Z} |t_n(z)|.$$

Again, if $\gamma < \alpha \leq \alpha_n^* < (\delta^2+\gamma^2)^{1/2}$, then

$$\max_{z\in E} t_n(z) = \min_{s_n\in S_n} \max_{z\in E} |s_n(z)|,$$

where α_n^* is close to $(\delta^2+\gamma^2)^{1/2}$ and $\lim_{n\to\infty} \alpha_n^* = (\delta^2+\gamma^2)^{1/2}$ [6], [7]. For n odd and $\gamma \leq \alpha < (\delta^2+\gamma^2)^{1/2}$, we have

$$\frac{T_n\left(\frac{\alpha}{\gamma}\right)}{\left[T_n^2\left(\left(1+\left(\frac{\delta}{\gamma}\right)^2\right)^{1/2}\right)-1\right]^{1/2}} = \max_{z\in E} |t_n(z)| \leq \frac{T_n\left(\frac{\alpha}{\gamma}\right)}{\left(T_n^2\left(\frac{\alpha}{\gamma}\right)-1\right)^{1/2}} \min_{s_n\in S_n} \max_{z\in E} |s_n(z)|.$$

Note that for $\alpha > \gamma$, we have $\lim_{n\to\infty} T_n(\alpha/\gamma)/(T_n^2(\alpha/\gamma)-1)^{1/2} = 1$, and thus the Chebychev polynomials are asymptotically optimal for n odd as well. We only focus on real foci or even n.

Using the definition $T_n(z) = \cosh(n\cosh^{-1}(z))$ and the formula $\cosh^{-1}(z) = \ln(z+(z^2-1)^{1/2})$, we may write

$$g(z) = z+\sqrt{z^2-1}, \qquad f(z) = z+\sqrt{z^2+1};$$

then

$$\frac{T_n\left(\frac{\alpha}{\gamma}\right)}{T_n\left(\frac{\delta}{\gamma}\right)} = \frac{g\left(\frac{\alpha}{\gamma}\right)^n + g\left(\frac{\alpha}{\gamma}\right)^{-n}}{g\left(\frac{\delta}{\gamma}\right)^n + g\left(\frac{\delta}{\gamma}\right)^{-n}}$$

and

$$\frac{T_n\left(\frac{\alpha}{\gamma}\right)}{T_n\left(\frac{(\delta^2+\gamma^2)^{1/2}}{\gamma}\right)} = \frac{g\left(\frac{\alpha}{\gamma}\right)^n + g\left(\frac{\alpha}{\gamma}\right)^{-n}}{f\left(\frac{\delta^2}{\gamma}\right)^n + f\left(\frac{\delta}{\gamma}\right)^{-n}}.$$

Thus, once δ, γ, and α are known, these bounds can be readily computed. These ideas have application to iterative methods for the solution of linear systems (see [10], [12]).

5. Bounding the diameter of an undirected graph from Chebychev polynomials. Let Q denote the Laplacian matrix of an undirected graph on n vertices. The following basic facts are summarized in [13].

Facts 5.1. Let G be an undirected graph. Then

(a) $Q(G)$ has only real eigenvalues and a complete set of orthonormal eigenvectors,

(b) $Q(G)$ is positive semidefinite,

(c) its smallest eigenvalue is $\lambda_1 = 0$, and a corresponding eigenvector is $u_1 = (1/\sqrt{n})(1, 1, \dots, 1)^*$. The multiplicity of 0 as an eigenvalue is equal to the number of components of G.

We assume that the eigenvalues of $Q(G)$ are ordered $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ in increasing order and repeated according to their multiplicity. Thus $\lambda_1 = 0$ and $\lambda_2 > 0$ if and only if G is connected. The following facts are known.

Facts 5.2. Let G have n vertices. Then

(a) $\lambda_2 \leq n/(n-1)\min\{d(v)|v \in V(G)\}$,

(b) $\lambda_n \leq \max\{d(u) + d(v)|uv \in E(G)\}$, and, if G is connected, equality holds if and only if G is bipartite and semiregular.

(c) $\lambda_n \leq n$ with equality if and only if the complement of G is not connected.

(d) $\sum_{i=1}^{n} \lambda_i = 2|E(G)| = \sum_{v\in V} d(v)$,

(e) $\lambda_n \geq n/(n-1)\max\{d(v)|v \in V(G)\}$.

(f) $\lambda_n \geq \max\{\sqrt{(d(v) - d(u))^2 + 4}|u, v \in V(G), u \neq v\}$.

THEOREM 5.3. *For an undirected graph G, we have*

$$D(G) \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}} \right\rfloor + 1.$$

Proof. Let

$$m \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}} \right\rfloor + 1.$$

Then

$$m > \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}}.$$

This can be rewritten as

$$T_m\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right) = \cosh\left(m\cosh^{-1}\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right) > n-1.$$

We know by our discussion in §4 that if we let

$$p_m(x) = \frac{T_m\left(\frac{\lambda_n+\lambda_2-2x}{\lambda_n-\lambda_2}\right)}{T_m\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right)},$$

then $p_m(0) = 1$, and

$$\max_{2 \le i \le n} |p_m(\lambda_i)| = \frac{1}{T_m\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right)} < \frac{1}{n-1}.$$

Now since Q is symmetric, we can write Q as

$$Q = \sum_{i=2}^{n} \lambda_i u_i u_i^*,$$

where the u_i are orthonormal. Note that u_1 is as defined in §3 and that $\lambda_1 = 0$. Thus

$$p_m(Q) - \frac{J}{n} = \sum_{i=2}^{n} p_m(\lambda_i) u_i u_i^*$$

and

$$\left\| p_m(Q) - \frac{J}{n} \right\| = \max_{2 \le i \le n} |p_m(\lambda_i)| < \frac{1}{n-1}.$$

Thus, by Theorem 3.2, we have $D \le m$, which completes the proof. □

Remark. Theorem 5.3 would be more asthetically pleasing if it yielded

$$D(G) \le \left\lceil \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right)} \right\rceil.$$

This inequality fails only if $D(G) = m + 1$ with

$$m = \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(\frac{\lambda_n+\lambda_2}{\lambda_n-\lambda_2}\right)}.$$

Does such a graph G exist? We easily see from the proof of Theorems 3.2 and 5.3 that for such a graph G, the Laplacian has eigenvector decomposition

$$Q = \sum_{i} \lambda_i u_i u_i^*,$$

where $0 = \lambda_1 < \lambda_2 \le \lambda_3 \cdots < \lambda_n$. There is a set of $m+1$ eigenvalues $\{\lambda_{k_i} | i = 1, \ldots, m+1\}$ such that $|p_m(\lambda_{k_i})| = 1/(n-1)$, where $p_m(\lambda)$ is the Chebychev polynomial defined in Theorem 5.3. That is,

$$\lambda_{k_i} = \frac{\lambda_n + \lambda_2}{2} - \frac{\lambda_n - \lambda_2}{2} \cos\left(\frac{(i-1)\pi}{m}\right) \quad \text{for } i = 1, \ldots, m+1.$$

Note that $\lambda_2 = \lambda_{k_1}, \lambda_n = \lambda_{k_{m+1}}$. There is a unique pair of indices $r \ne s$ such that $\langle e_r, Q^j e_s \rangle = 0$ for $i = 0, \ldots, m$. Furthermore,

$$e_r = \frac{1}{\sqrt{n}} u_1 + \sum_{i=1}^{m+1} \alpha_j u_{k_i},$$

$$e_s = \frac{1}{\sqrt{n}} u_1 + \sum_{i=1}^{m+1} \alpha_i (-1)^i u_{k_i},$$

where

$$\alpha_i^2 = \frac{(-1)^{i-1}}{n} \frac{\prod_{j \neq i} \lambda_{k_i}}{\prod_{j \neq i} \left(\lambda_{k_j} - \lambda_{k_i}\right)}.$$

Finally, we have, for $j = 1, \ldots, m$,

$$\sum_{i \,\mathrm{odd}} \alpha_i^2 \lambda_{k_i}^j = \sum_{i \,\mathrm{even}} \alpha_i^2 \lambda_{k_i}^j = \langle e_r, Q^j, e_r \rangle = \langle e_s, Q^j e_s \rangle ,$$

which implies that $d(r) = d(s)$, while

$$\sum_{i \,\mathrm{odd}} \alpha_i^2 = \frac{1}{2}, \qquad \sum_{i \,\mathrm{even}} \alpha_i^2 = \frac{1}{2} - \frac{1}{n}.$$

It is not known if graphs satisfying these conditions exist.

The following corollary is useful in conjunction with the estimates given in Facts 5.2.

COROLLARY 5.4. *Let s_2 and s_n be estimates of λ_2 and λ_n, respectively, such that*

$$0 < s_2 \leq \lambda_2 \leq \lambda_n \leq s_n.$$

Then

$$D(G) \leq \left\lfloor \frac{\cosh -1 \,(n-1)}{\cosh^{-1} \frac{s_n + s_2}{s_n - s_2}} \right\rfloor + 1.$$

Proof. It is easy to see that

$$\cosh^{-1} \frac{\lambda_n + \lambda_2}{\lambda_n - \lambda_2} \geq \cosh^{-1} \frac{s_n + s_2}{s_n - s_2}. \qquad \square$$

COROLLARY 5.5. *Let A denote the adjacency matrix of a k-regular graph G. Suppose that A has second largest eigenvalue $\tilde{\lambda}$ (in absolute value). Then*

$$D(G) \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(k/\tilde{\lambda}\right)} \right\rfloor + 1.$$

Proof. It is easy to see that

$$k - \tilde{\lambda} \leq \lambda_2 \leq \lambda_n \leq k + \tilde{\lambda}.$$

Therefore

$$D(G) \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1} \frac{\lambda_n + \lambda_2}{\lambda_n - \lambda_2}} \right\rfloor + 1 \leq \left\lfloor \frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(k/\tilde{\lambda}\right)} \right\rfloor + 1.$$

We remark that

$$\frac{\cosh^{-1}(n-1)}{\cosh^{-1}\left(k/\tilde{\lambda}\right)} < \frac{\ln(n-1)}{\ln\left(k/\tilde{\lambda}\right)}$$

except for the complete graph; thus Theorem 5.3 yields an improvement over the bound $D(G) \leq \lceil \ln(n-1)/\ln(k/\tilde{\lambda}) \rceil$ in [3]. $\square$

Remark. Chebychev polynomials have also been used to bound the diameter of Ramanujan graphs by Lubotzky, Phillips, and Sarnak [9] and for general graphs by Sarnak [16].

6. General directed graphs.

In this section, we examine the estimates we can make for the diameter of general directed graphs G. Our starting point is Theorem 3.2, and, as in that section, we assume that the in-degree of each vertex of G equals its out-degree. (This holds if and only if G is Eulerian.) Again, $Q(G)$ is the Laplacian of G.

THEOREM 6.1. *If ρ and λ are such that $1 \geq \rho > ||I - \lambda Q - J/n||$, then*

$$D(G) \leq \left\lceil \frac{\ln(n-1)}{\ln \frac{1}{\rho}} \right\rceil .$$

Proof. Let

$$m = \left\lceil \frac{\ln(n-1)}{\ln\left(\frac{1}{\rho}\right)} \right\rceil .$$

So,

$$m \geq \frac{\ln(n-1)}{\ln\left(\frac{1}{\rho}\right)}.$$

Let $p_m(x) = (1 - \lambda x)^m$. Then

$$p_m(Q) - \frac{J}{n} = \left(I - \lambda Q - \frac{J}{n}\right)^m ,$$

so

$$\begin{aligned} \left\| p_m(Q) - \frac{J}{n} \right\| &= \left\| \left(I - \lambda Q - \frac{J}{n}\right)^m \right\| \\ &\leq \left\| I - \lambda Q - \frac{J}{n} \right\|^m \\ &< \rho^m = e^{-m \ln(1/\rho)} \\ &\leq e^{-\ln(n-1)} = \frac{1}{n-1}. \end{aligned}$$

Thus, by Theorem 3.2, $D \leq m$. This proves the theorem. □

From this theorem, we see that it would be useful to know that

$$\Phi = \min_{\lambda} \left\| I - \lambda Q - \frac{J}{n} \right\| .$$

We can give an estimate of Φ to show that this minimum is less than 1 when G is connected. If B is a matrix, we let $B_S = (B + B^T)/2$ be the symmetric part of B and we let $B_A = (B - B^T)/2$ be the antisymmetric part of B. If G is connected, then Q_S is the Laplacian of a weighted connected undirected graph (see [13]), so $\sigma = \lambda_2(Q_S) > 0$. We let $\mu = ||Q||$. In the proof of Theorem 6.3, we show that $\Phi \leq \sqrt{\mu^2 - \sigma^2}/\mu$, so $\Phi < 1$ when G is connected. First, we need a technical lemma.

LEMMA 6.2. *Let μ^2 be the largest eigenvalue of $B^T B$ and $\sigma \neq \mu$ be the smallest eigenvalue of B_S. If there is a unit vector u such that*

$$\sigma u = B_S u,$$

$$\mu^2 u = B^T B u,$$

then there is a unit vector v orthogonal to u such that the span S of $\{u, v\}$ is an invariant subspace of B and the matrix of B relative to $\{u, v\}$ is given by

$$M = \begin{pmatrix} \sigma & -\sqrt{\mu^2 - \sigma^2} \\ \sqrt{\mu^2 - \sigma^2} & \sigma \end{pmatrix}.$$

Proof. First, we show that

$$(B_S B_A - B_A B_S)\, u = 0,$$

$$B_A^2 u = (\sigma^2 - \mu^2)\, u.$$

We repeatedly use the fact that B_A antisymmetric implies that $(B_A x, x) = 0$. We have

$$\begin{aligned} \mu^2 u = B^T B u &= (B_S - B_A)\,(B_S + B_A)\, u \\ &= (B_S^2 - B_A^2)\, u + (B_S B_A - B_A B_S)\, u \\ &= (\sigma^2 I - B_A^2)\, u + (B_S - \sigma I)\, B_A u. \end{aligned}$$

Since $(B_A u, \mu^2 u) = 0$,

$$\begin{aligned} 0 &= (B_A u, (\sigma^2 I - B_A^2)\, u + (B_S - \sigma I)\, B_A u) \\ &= (B_A u, (B_S - \sigma I)\, B_A u)\,. \end{aligned}$$

Since σ is the smallest eigenvalue of B_S, the square root of $B_S - \sigma I$ exists. Thus

$$0 = \left(\sqrt{B_S - \sigma I}\,(B_A u)\,, \sqrt{B_S - \sigma I}\,(B_A u)\right),$$

so $(B_S - \sigma I) B_A u = 0$. This yields both of the claimed identities.

Now let $\alpha = ||B_A u|| = \sqrt{(B_A u, B_A u)} = \sqrt{\mu^2 - \sigma^2}$ and $v = (1/\alpha) B_A u$. Then $(v, u) = (B_A u, u) = 0$. In addition,

$$B_S v = \frac{1}{\alpha}\, B_S B_A u = \frac{1}{\alpha}\, B_A B_S u = \sigma v,$$

while

$$B_A v = \frac{1}{\alpha}\, B_A^2 u = \frac{\sigma^2 - \mu^2}{\alpha}\, u = -\alpha u.$$

This proves the lemma. □

THEOREM 6.3. *If G is connected and not complete or a directed 3-cycle,*

$$D\,(G) \leq \left\lceil \frac{\ln\,(n-1)}{\ln \frac{\mu}{\sqrt{\mu^2 - \sigma^2}}} \right\rceil,$$

where σ is the smallest nonzero eigenvalue of Q_S and $\mu = ||Q||$.

Proof. Note that for a complete graph, $\mu = \sigma$, and for the directed 3-cycle, $\mu = \sqrt{3}$, while $\sigma = \frac{3}{2}$. Thus the inequality fails for these cases.

For any λ, we have

$$\left\| I - \lambda Q - \frac{J}{n} \right\|^2 = \max_{w \neq 0} \frac{\left(\left(I - \lambda Q - \frac{J}{n}\right) w, \left(I - \lambda Q - \frac{J}{n}\right) w\right)}{(w, w)}.$$

Let $w^{\perp} = w - (w, u_1)u_1$, where $u_1 = (1/\sqrt{n})(1, 1, \ldots, 1)^*$. Then $(w^{\perp}, u_1) = 0$. Since $(I - \lambda Q - (J/n))u_1 = 0$ and $(J/n)w^{\perp} = 0$, we have

$$\begin{aligned} \frac{\left(\left(I - \lambda Q - \frac{J}{n}\right) w, \left(I - \lambda Q - \frac{J}{n}\right) w\right)}{(w, w)} &= \frac{\left((I - \lambda Q)\, w^{\perp}, (I - \lambda Q)\, w^{\perp}\right)}{(w^{\perp}, w^{\perp}) + (w, u_1)^2} \\ &\leq \frac{\left((I - 2\lambda Q_S + \lambda^2 Q^T Q)\, w^{\perp}, w^{\perp}\right)}{(w^{\perp}, w^{\perp})} \\ &\leq 1 - 2\lambda\sigma + \lambda^2\mu^2. \end{aligned}$$

Setting $\lambda = \sigma/\mu^2$ yields

$$\left\| I - \lambda Q - \frac{J}{n} \right\|^2 \leq 1 - \frac{\sigma^2}{\mu^2} < 1,$$

since G is connected. Thus

$$\left\| I - \lambda Q - \frac{J}{n} \right\| = \frac{\sqrt{\mu^2 - \sigma^2}}{\mu} < 1.$$

If this is a strict inequality, then the theorem follows from Theorem 6.1. Equality can hold only if there is a vector u such that $Q_S u = \sigma u$ and $Q^T Q u = \mu^2 u$. Let $C = I - \lambda Q$. Let

$$m = \frac{\ln(n-1)}{\ln \frac{\mu}{\sqrt{\mu^2 - \sigma^2}}}.$$

If we examine the proofs of Theorems 6.1 and 3.2, we find that $D(G) \leq \lfloor m \rfloor + 1$, and that $D(G) > \lceil m \rceil$ can hold only if m is an integer and if there exists $r \neq s$ such that

$$-\frac{1}{n} = \left(e_r, \left(C^k - \frac{J}{n}\right) e_s\right) = \left(f_r, \left(C^k - \frac{J}{n}\right) f_s\right) = (f_r, C^k f_s)$$

for $1 \leq k \leq m$ and $||C - (J/n)|| = \sqrt{\mu^2 - \sigma^2}/\mu$. Thus we can take $u = f_s/||f_s|| = \sqrt{n/(n-1)}f_s$ and $B = \tilde{Q} = (Q$ restricted to the space orthogonal to $u_1)$ in Lemma 6.2 and $v = 1/\sqrt{\mu^2 - \sigma^2} B_A u$. The action of C on $S = \{u, v\}$ is then given by

$$M = \begin{pmatrix} 1 - \lambda\sigma & \lambda\alpha \\ -\lambda\alpha & 1 - \lambda\sigma \end{pmatrix},$$

and the action of C^2 is given by

$$M^2 = \begin{pmatrix} (1 - \lambda\sigma)^2 - (\lambda\alpha)^2 & 2\lambda\alpha\,(1 - \lambda\sigma) \\ -2\lambda\alpha\,(1 - \lambda\sigma) & (1 - \lambda\sigma)^2 - (\lambda\alpha)^2 \end{pmatrix}.$$

This yields the system

$$(f_r, Cu) = (1 - \lambda\sigma)\,(f_r, u) - \lambda\alpha\,(f_r, v)\,,$$

$$(f_r, C^2 u) = \left[(1 - \lambda\sigma)^2 - (\lambda\alpha)^2\right](f_r, u) - 2\lambda\alpha\,(1 - \lambda\sigma)\,(f_r, v)\,.$$

If we multiply these equations by $\sqrt{(n-1)/n}$, we obtain

$$(f_r, Cf_s) = (1-\lambda\sigma)(f_r, f_s) - \lambda\alpha\sqrt{\frac{n-1}{n}}(f_r, v),$$

$$(f_r, C^2 f_s) = \left[(1-\lambda\sigma)^2 - (\lambda\alpha)^2\right](f_r, f_s) - 2\lambda\alpha(1-\lambda\sigma)\sqrt{\frac{n-1}{n}}(f_r, v).$$

If $m \geq 2$ and we use $(f_r, f_s) = -1/n$, we obtain

$$-\frac{1}{n} = -\frac{1}{n}(1-\lambda\sigma) - \lambda\alpha\sqrt{\frac{n-1}{n}}(f_r, v),$$

$$-\frac{1}{n} = -\frac{1}{n}\left[(1-\lambda\sigma)^2 - (\lambda\alpha)^2\right] - 2\lambda\alpha(1-\lambda\sigma)\sqrt{\frac{n-1}{n}}(f_r, v).$$

Elimination yields

$$\frac{\sigma^2}{\mu^4}(\mu^2 - \sigma^2) = (\lambda\alpha)^2 = 0,$$

which has no roots but $\sigma = 0$ or $\mu^2 = \sigma^2$.

Thus the only possibilities are $\sigma = \mu$ or $m = 1$. If $\sigma = \mu$, then all the eigenvalues of B_S and all the singular values of B are equal to μ. Thus every vector u satisfies the conditions of Lemma 6.2. In particular, $B_A^2 = 0$, so $B = B_S = \mu I$ and $Q = \mu(I - (J/n))$. Since the entries of Q must be integers, $\mu = n$. Thus the graph is the complete graph.

If $m = 1$, then $\mu^2/(\mu^2 - \sigma^2) = (n-1)^2$, and so $||I - \lambda Q - (J/n)|| = 1/(n-1)$. Because of $(f_r, (I - \lambda Q - (J/n))f_s) = -1/n$, we have, in this case,

$$\left(I - \lambda Q - \frac{J}{n}\right) f_s = -\frac{1}{n-1} f_r.$$

This leads to

$$\frac{\sigma}{\mu^2} Qe_s = \lambda Q e_s = e_s + \frac{1}{n-1} e_r - \frac{1}{n-1} w,$$

and so

$$(e_s, Qe_s) = \frac{\mu^2}{\sigma}\frac{n-2}{n-1}.$$

If $n > 2$, then there is a $t \neq r, s$, and

$$(e_t, Qe_s) = -\frac{\mu^2}{\sigma}\frac{1}{n-1}.$$

Since the off-diagonal entries of Q are 0 or -1, we must have $\mu^2/\sigma = n-1$. Thus $(e_s, Qe_s) = n-2$. We may also solve for both μ and σ to obtain $\mu = \sqrt{n(n-2)}$ and $\sigma = n(n-2)/(n-1)$. Let prime denote parameters relative to the complement of G. We have

$$Q'_S + Q_S = n\left(I - \frac{J}{n}\right).$$

Also, note that

$$\sigma_{\max}(Q'_S) = \max_{||x||=1}(Qx, x) \geq \max_i(e_i, Qe_i) = \max_i d_i(G').$$

Thus

$$\frac{n(n-2)}{n-1} = \sigma = n - \sigma_{\max}(Q'_S) \le n - \max_i d_i(Q'_S),$$

so if $n > 2$,

$$\max_i d_i(Q'_S) \le 1 + \frac{1}{n-1} < 2.$$

Thus each vertex of G' has degree 0 or 1, so G' is the disjoint union of isolated vertices and directed cycles D_{k_i} with $\sum k_i \le n$. We need

$$\begin{aligned} 1 + \frac{1}{n-1} &= n - \sigma = \sigma_{\max}(Q'_S) = \max_i \sigma_{\max}(Q(D_{K_i}))_S \\ &= \frac{1}{2} \max_i \sigma_{\max}(Q(C_{k_i})) \\ &= \begin{cases} 2 & \text{if some } k_i \text{ is even,} \\ 1 - \cos\frac{2\pi}{k}, k = \max_i k_i & \text{if no } k_i \text{ is even.} \end{cases} \end{aligned}$$

However, we can never have $-\cos(2\pi/k) = 1/(n-1)$ unless $k = 3$ and $n = 3$; that is, G is a directed 3-cycle. This contradiction eliminates this case and completes the proof of the theorem. □

7. Normal matrices. If we compare the result of Theorem 6.3 with those obtained from Theorem 5.3 for undirected graphs, we see a marked difference. We would like to extend the method of §5 to directed graphs. We can do this fairly easily for one class of directed graphs, namely, those whose adjacency matrix is normal. A matrix A is normal if $AA^* = A^*A$. We have the following well-known fact. (For the application of normal matrices to iterative methods, see [5], which also includes some other important properties of normal matrices.)

LEMMA 7.1. *The following properties of a matrix A are equivalent*:

(i) $A^*A = AA^*$,

(ii) *A^* is a polynomial in A*,

(iii) *the eigenvectors of A form a complete orthonormal basis.*

THEOREM 7.2. *Suppose that a directed graph with n vertices has Laplacian Q such that $QQ^T = Q^TQ$ and that all the eigenvalues of Q except 0 are contained in an ellipse with center $\delta > 0$, foci $\delta \pm \gamma, \gamma > 0$, and semi-major axis of length α, where $\gamma < \alpha < \delta$. If m is the least integer such that*

$$\frac{T_m\left(\frac{\alpha}{\gamma}\right)}{T_m\left(\frac{\alpha}{\delta}\right)} < \frac{1}{n-1},$$

then $D(G) \le m$. Furthermore, suppose that all of the eigenvalues of Q except 0 are contained in an ellipse with center $\delta > 0$, foci $\delta \pm i\gamma, \gamma > 0$, and semi-major axis of length $\alpha, \gamma < \alpha < \sqrt{\delta^2 + \gamma^2}$. If m is the least even integer such that

$$\frac{T_m\left(\frac{\alpha}{\gamma}\right)}{T_m\left(\sqrt{1 + \left(\frac{\delta}{\gamma}\right)^2}\right)} < \frac{1}{n-1},$$

then $D(G) \le m$.

Proof. The proof relies on the remarks of §4 and is similar to the proof of Theorem 5.3. We leave it to the reader. □

The following lemma may be of some use in estimating the diameter of directed graphs in conjunction with Theorem 7.2.

LEMMA 7.3. *Let Q be the Laplacian of a connected directed graph with in-degree equal to out-degree. Then the eigenvalues of Q lie in a disk of radius $\Delta = \max\{d(u) | u \in G\}$ with center at Δ. There is exactly one zero eigenvalue. Furthermore, if there are $k > 1$ eigenvalues on the boundary of the disk $|z - \Delta| = \Delta$, then these eigenvalues have the form $(1 - e^{(2\pi m/k)i})\Delta$, and G is k-partite.*

Proof. That the eigenvalues λ satisfy $|\lambda - \Delta| \leq \Delta$ follows from the Gershgorin theorem. Now consider the nonnegative matrix $B = \Delta I - Q$. The Perron–Frobenius theorem yields the result. □

8. Estimating the ellipse which encloses the eigenvalues of a real normal matrix.

It is easy to recognize when Theorem 7.2 applies. For example, if Γ is a group and $\Delta = \{\delta_1, \delta_2, \dots, \delta_d\}$ are generators, the Cayley graph $G(\Gamma, \Delta)$ has a normal adjacency matrix if and only if, for each $g \in \Gamma$, the number of solutions (δ_i, δ_j) to $\delta_i g = \delta_j$ is the same as the number of solutions (δ_i, δ_j) to $g\delta_i = \delta_j$. Thus, if Γ is abelian, Q is normal. We need a method for finding the ellipse in Theorem 7.2.

Suppose that the eigenvalues of Q are known. There are an infinite number of ellipses symmetric with respect to the real axis that enclose the nonzero eigenvalues of Q and exclude zero. We would like to choose the ellipse that minimizes m in Theorem 7.2. Equivalently, given any m, we would like to find the ellipse that minimizes the bounds in Theorem 7.2. To our knowledge, this problem is unsolved. However, a related problem has been solved and may be useful. Using the definition $T_m(z) = \cosh(n \cosh^{-1}(z))$ and making use of the formula $\cosh^{-1}(z) = \ln(z + (z^2 - 1)^{1/2})$, we may write $g(z) = z + \sqrt{z^2 - 1}, f(z) = z + \sqrt{z^2 + 1}$, and

$$\frac{T_m\left(\frac{\alpha}{\gamma}\right)}{T_m\left(\frac{\delta}{\gamma}\right)} = \left(\frac{\alpha + \sqrt{\alpha^2 - \gamma^2}}{\delta + \sqrt{\delta^2 + \gamma^2}}\right)^m \left(\frac{1 + g\left(\frac{\alpha}{\gamma}\right)^{-2m}}{1 + g\left(\frac{\delta}{\gamma}\right)^{-2m}}\right)$$

for the case of real foci, and

$$\frac{T_m\left(\frac{\alpha}{\gamma}\right)}{T_m\left(\frac{\sqrt{\delta^2+\gamma^2}}{\gamma}\right)} = \left(\frac{\alpha + \sqrt{\alpha^2 - \gamma^2}}{\delta + \sqrt{\delta^2 + \gamma^2}}\right)^m \left(\frac{1 + g\left(\frac{\alpha}{\gamma}\right)^{-2m}}{1 + f\left(\frac{\delta}{\gamma}\right)^{-2m}}\right)$$

for complex foci and m even.

The second term in each expression is bounded by 2 and rapidly approaches L for large m. The problem of finding the ellipse symmetric with respect to the real axis that encloses a given set of points and excludes zero that minimizes the first term in the above expression is solved in Manteuffel [10]. The solution depends only on the convex hull of the nonzero spectrum of Q. If the best such ellipse has real foci, then the smallest m that satisfies

$$\frac{T_m\left(\frac{\alpha}{\gamma}\right)}{T_m\left(\frac{\delta}{\gamma}\right)} < \frac{1}{n-1}$$

can be found. If the best such ellipse has complex foci, then the smallest even m that satisfies

$$\frac{T_m\left(\frac{\alpha}{\gamma}\right)}{T_m\left(\frac{(\delta^2+\gamma^2)^{1/2}}{\gamma}\right)} < \frac{1}{n-1}$$

can be found. Application of Theorem 7.2 then yields the bound $D(G) \leq m$.

9. Estimating the spectrum of Q. The theorems in the above sections require knowledge of the extremal eigenvalues of Q. If the eigenvalues of Q are not known, they can be computed by a variety of methods. If the dimension of Q is not too large, good algorithms exist for computing the entire spectrum. If the dimension is large, it may be more practical to estimate the extremal eigenvalues by iterative methods. If systems of the type $(\lambda I - Q)x = y$ are easily solved, then inverse iteration may be employed (cf. Wilkinson [17]). If not, then methods based upon taking orthogonal sections of Q may be used, for example, the Lanczos algorithm (cf. [8]). These methods only require repeated matrix vector multiplication.

Suppose that Q is symmetric. The Lanczos algorithm requires an initial vector, say r, and iteratively forms a tridiagonal matrix T_k that is the orthogonal section of Q onto the Krylov subspace $K_k(r, Q) = sp\{r, Qr, \ldots, Q^{k-1}r\}$. The eigenvalues of T_k yield good approximations to the extremal eigenvalues of Q. If r is chosen to be orthogonal to the null vector u_1, then the extremal eigenvalues of T_k approximate λ_2 and λ_n. Estimates of the rate of convergence are available [4].

If Q is nonsymmetric then Arnoldi's method [15], [2] yields an upper Hessenberg matrix H_k instead of T_k as the orthogonal section of Q onto the Krylov space $K_k(r, Q)$. The eigenvalues of H_k again yield approximations to the extremal eigenvalues of Q. In this case, however, the convex hull of the spectrum of H_k approximates the convex hull of the spectrum of Q. Again, if r is chosen orthogonal to u_1, the approximations ignore the zero eigenvalue. If Q is normal, the only case in which we seek the spectrum of a nonsymmetric Q, the procedure yields good approximations. The adaptive methods described in [11] simultaneously find the eigenvalues and the optimal ellipse.

For general Q, we are limited to Theorem 6.2. This requires $\mu = ||Q||$, which can be computed by applying the Lanczos procedure on Q^TQ to find $\mu^2 = ||Q^TQ||$. The value $\sigma = \lambda_2(Q_S)$ can also be approximated by the Lanczos procedure applied to Q_S.

REFERENCES

[1] N. Alon and V. D. Milman, λ_1, *Isoperimetric inequalities for graphs and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[2] W. E. Arnoldi, *The principal of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[3] F. R. K. Chung, *Diameters and eigenvalues*, J. Amer. Math. Soc., 2 (1989), pp. 187–196.

[4] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[5] V. Faber and T. A. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[6] B. Fischer and P. Freund, *On the constrained Chebychev approximation problem on ellipses*, J. Approx. Theory, 62 (1990), pp. 297–315.

[7] ———, *Chebychev polynomials are not always optimal*, J. Approx. Theory, 65 (1991), pp. 261–272.

[8] G. A. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, MD, 1983.

[9] A. Lubotzky, R. Phillips, and P. Sarnak, *Ramanujan graphs*, Combinatorica, 8 (1988), pp. 261–277.

[10] T. A. Manteuffel, *The Tchebychev iteration for nonsymmetric systems*, Numer. Math., 28 (1977), pp. 307–327.

[11] ———, *The adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*, Numer. Math., 31 (1978), pp. 183–208.

[12] B. Mohar, *Eigenvalues, diameter, and mean distance in graphs*, Graphs Combin., 7 (1991), pp. 53–64.

[13] ———, *The Laplacian spectrum of graphs*, in Graph Theory, Combinatorics and Applications 2 (Kalamazoo, MI, 1988), Wiley-Intersciences, pp. 871–898.

[14] T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley, New York, 1990.

[15] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[16] P. Sarnak, *Some Applications of Modular Forms*, Cambridge University Press, Cambridge, UK, 1990.

[17] J. H. Wilkinson, *Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 458–473, August 1994

RANDOM RESOURCE ALLOCATION GRAPHS AND THE PROBABILITY OF DEADLOCK*

JAMES F. LYNCH†

Abstract. Resource allocation graphs are directed bipartite graphs satisfying certain constraints. The two partitions of the vertex set correspond to the processes and resources of a multiprogramming computer system, and the edges indicate resource allocations and requests. When all the resources are distinguishable, existence of a cycle in such a graph is equivalent to deadlock in the system. The main results here are exact and asymptotic formulas for the number of resource allocation graphs and acyclic resource allocation graphs with m processes, n allocated resources, and q edges. If a uniform probability distribution is taken on such graphs, then the proportion of blocked processes, $(q-n)/m$, determines the asymptotic behavior of the probability of deadlock: If $(q-n)/m \to 0$, then the probability of deadlock is asymptotic to 0, and if $(q-n)/m \to \alpha > 0$, then the probability is asymptotic to a positive value. Similar formulas are also given in terms of m and n only.

Key words. operating systems, resource allocation, random graphs, deadlock

AMS subject classifications. 68R10, 05C80

1. Introduction.

Resource allocation graphs are directed bipartite graphs that describe the state of a multiprocessing computer system. The two partitions of the vertex set, P and R, are called the *processes* and the *resources*, respectively. In this paper we assume that all resources are distinguishable; a more general model would allow several instances of a resource type. An edge from a resource to a process indicates that the resource is *allocated* to the process. An edge from a process to a resource indicates that the process has requested but not yet received the resource, i.e., it is *blocked*. This occurs when the resource is already allocated to another process. For example, in Fig. 1, r_1 is allocated to p_2 and r_2 is allocated to p_1, but p_1 is blocked. We assume that processes request resources one at a time. Again, this restriction could be relaxed.

With this interpretation, it is easily seen that a resource allocation graph must satisfy the following conditions:

1. Every process has outdegree at most 1. (Once a process is blocked, it cannot make any more requests.)
2. Every resource has outdegree at most 1. (A resource cannot be shared.)
3. If (p,r) is an edge of the graph for some $p \in P$ and $r \in R$, then there must be an edge (r,q) for some $q \neq p$. (If a resource is requested by a process and it is not already allocated, then the request must be granted.)

Conversely, any directed bipartite graph satisfying these conditions corresponds to a possible state of a multiprocessing system. Thus we take as our definition of resource allocation graph a triple $\langle P, R, E\rangle$, where P and R are disjoint sets and $E \subset P \times R \cup R \times P$ is a set of edges satisfying conditions 1–3.

A *cycle* in a resource allocation graph is a sequence $p_1, r_1, p_2, r_2, \ldots, p_k, r_k$, where $\{p_1, \ldots, p_k\} \subseteq P$, $\{r_1, \ldots, r_k\} \subseteq R$, and $\{(p_1,r_1),(r_1,p_2),(p_2,r_2),\ldots,(p_k,r_k),(r_k,p_1)\} \subseteq E$. When there is a single instance of each resource type, existence of

* Received by the editors April 20, 1992; accepted for publication (in revised form) April 12, 1993. This research was supported by National Science Foundation grants CCR-8805880 and CCR-9006303.

† Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13699-5815 (jlynch@sun.mcs.clarkson.edu).

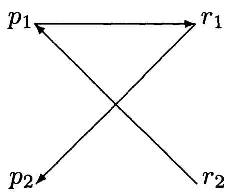


FIG. 1

a cycle is equivalent to the condition of *deadlock*: the processes in the cycle are blocking each other indefinitely. In the example of Fig. 1, if p_2 requests r_2, then p_1 and p_2 will be deadlocked.

Deadlock is a situation of considerable practical importance, and it is expected that it will become even more pervasive as parallel computation and dynamic resource allocation become more prevalent. Hofri [6] has written an annotated bibliography that lists over 150 papers on deadlock, most of them describing ways of dealing with deadlock. The two main approaches are: ensuring that it does not occur, or detecting and recovering from it when it does occur. Either scheme is costly to implement, as are the many variations that have been proposed.

Because of the difficulty of knowing when deadlock occurs, it would be desirable to have some means of estimating the likelihood or frequency of deadlock. This could help in determining the method used to handle deadlock, for example, deciding how often a deadlock detection algorithm should be invoked. The results of this paper show that the probability of deadlock is approximated by a simple function of certain easily measurable parameters of the system. Of course, this depends on the probability distribution of random resource allocation graphs. Since our aims are rather general and theoretical, we assume all resources have equal probabilities of being requested or allocated. Nevertheless, we believe that the techniques can be extended to more complex models of random resource allocation graphs, where the probability distributions need not be uniform.

The theory of random graphs, initiated by Erdös and Rényi [3], suggests several models for random structures. The most widely used, the independent edge probability model, is not appropriate because the edges in a resource allocation graph are not independent of each other (see conditions 1–3, above). A better model is obtained by considering a uniform distribution on all resource allocation graphs with a given number of processes, resources, and edges, i.e., resource requests either granted but not yet released, or not yet granted. However, this allows for the possibility of unallocated, i.e., isolated, resources. As will be seen, this complicates the analysis of the asymptotics.

The more tractable model that we concentrate on uses a uniform distribution on resource allocation graphs with m processes, n allocated resources, and q edges. The intent is that these three parameters characterize the "load" on the system, or its level of activity. Also, they are easily measurable in actual systems.

Our main results are exact and asymptotic formulas for the number of acyclic, i.e., deadlock free, resource allocation graphs, and the number of resource allocation graphs, in terms of m, n, and q. They imply, not surprisingly, that the proportion of blocked processes $(q-n)/m$ determines the asymptotic behavior of the probability of deadlock. That is, if $(q-n)/m \to 0$, then the probability of deadlock is asymptotic to 0, and if $(q-n)/m \to \alpha > 0$, then the probability is asymptotic to a positive value.

Further, with certain constraints on the growth rates of m, n, and q, the probability of deadlock is asymptotic to $(q-n)^2/2m^2$. Then, if the following assumptions are made, the expected time until deadlock is $\Theta(m^2/(q-n)^2)$. First, the system is a Markov chain. That is, the future evolution of the resource allocation graph depends only on its current configuration. Second, after a brief time, the probability distribution is close to its limiting distribution. That is, the information about its initial configuration is quickly lost. This is similar to many examples in statistical mechanics, where systems of interacting particles are studied. If the transition probabilities are too complex or not even known, then some limiting distribution is assumed. Even with such simplifications, there can be close agreement with observed macroscopic phenomena, and useful predictions can be made.

There are obvious analogies to the theory of random graphs. In particular, in [3] it was shown that for a random graph with n vertices and q edges, the asymptotic probability of existence of a cycle is 0 if and only if the asymptotic value of q/n is 0. In particular, if $q/n \geq 1/2$, then the asymptotic probability is 1. Here, of course, the rules of evolution are quite simple: at each step, an edge is added with a uniform probability distribution.

2. Main results.

In this section, we state and prove results about resource allocation graphs $\langle P, R, E\rangle$ with a given number of processes, allocated resources, and edges. We assume that there are no free resources. In §3, we derive analogous results for resource allocation graphs with a given number of processes and resources. In that case, we are able to find simple asymptotic expressions even when free resources are allowed.

For natural numbers m, n, and q, let

$$\begin{aligned} a(m,n,q) &= \text{the number of acyclic resource allocation graphs with} \\ &\quad |P| = m,\ |R| = n, \text{ and } |E| = q; \\ g(m,n,q) &= \text{the number of resource allocation graphs with} \\ &\quad |P| = m,\ |R| = n, \text{ and } |E| = q; \text{ and} \\ p(m,n,q) &= 1 - \frac{a(m,n,q)}{g(m,n,q)}, \text{ the probability of deadlock for a randomly selected} \\ &\quad \text{resource allocation graph with } |P| = m,\ |R| = n, \text{ and } |E| = q. \end{aligned}$$

The proofs of the exact formulas for these functions use generating functions and formal power series techniques. A thorough exposition of these methods can be found in Goulden and Jackson [5]. Austin [1] and Scoins [7] used a similar approach to count the number of $m \times n$ bipartite trees.

It will be helpful to define some special kinds of directed bipartite graphs. A *bipartite tree* is an acyclic, connected bipartite graph. For every bipartite tree, there is a unique vertex of outdegree 0, which is called the *root.* All the other vertices are directed toward the root. If the root is in P, the tree is said to be a *P-tree*; otherwise it is an *R-tree.* A *P-forest* (respectively, *R-forest*) is a set of P-trees (respectively, R-trees). A *bipartite cycle* is a connected bipartite graph such that every vertex has indegree 1 and outdegree 1.

Let

$$t(m,n,q) = \text{the number of } P\text{-trees with } |P| = m,\ |R| = n, \text{ and } |E| = q;$$

$u(m,n,q) =$ the number of R-trees with $|P| = m$, $|R| = n$, and $|E| = q$; and
$c(m,n,q) =$ the number of bipartite cycles with $|P| = m$, $|R| = n$, and $|E| = q$.

Let

$$G(x,y,z) = \sum_{m,n,q} \frac{g(m,n,q)}{m!n!} x^m y^n z^q$$

and $T(x,y,z)$, $U(x,y,z)$, and $C(x,y,z)$ be the corresponding generating functions for t, u, and c, respectively. The generating functions are exponential with respect to processes and resources because they are labeled, but ordinary with respect to edges because they are not.

LEMMA 2.1. *We have*

$$T(x,y,z) = xe^{zU(x,y,z)}, \tag{2.1}$$

$$U(x,y,z) = ye^{zT(x,y,z)}. \tag{2.2}$$

Proof. We prove only (2.1); the proof of (2.2) is symmetric. We use a variation of the labeled branch decomposition for trees ([5, §3.3.9]). Every P-tree can be decomposed into the root r, j R-trees for some $j \geq 0$, and j edges from the roots of the R-trees to r. For fixed j, the generating function for this set is

$$\frac{xU(x,y,z)^j z^j}{j!}.$$

Summing over j gives us (2.1). □

LEMMA 2.2. *We have*

$$C(x,y,z) = \log\left(\left(1 - xyz^2\right)^{-1}\right).$$

Proof. It is evident that

$$c(m,n,q) = \begin{cases} (m-1)!m! & \text{if } n = m \text{ and } q = 2m, \\ 0 & \text{otherwise.} \end{cases}$$

The lemma follows immediately. □

LEMMA 2.3. *We have*

$$G(x,y,z) = \left(1 - T(x,y,z)U(x,y,z)z^2\right)^{-1} \exp\left(T(x,y,z) - T(x,y,z)U(x,y,z)z^2\right).$$

Proof. Every connected resource allocation graph is either a P-tree (as in Fig. 2(a)), or it consists of a bipartite cycle of length at least 4 incident with bipartite trees such that the root of each tree is the only vertex in common with the cycle (as in Fig. 2(b)), or it is an isolated resource. Since we are counting only allocated resources, the third case can be ignored. The generating function for the set of connected resource allocation graphs containing a bipartite cycle of length $2m$ is

$$\frac{c(m,m,2m)T(x,y,z)^m U(x,y,z)^m z^{2m}}{(m!)^2}.$$

Summing over m, $C(T(x,y,z), U(x,y,z), z) - T(x,y,z)U(x,y,z)z^2$ is the generating function for the set of connected resource allocation graphs with a cycle of length

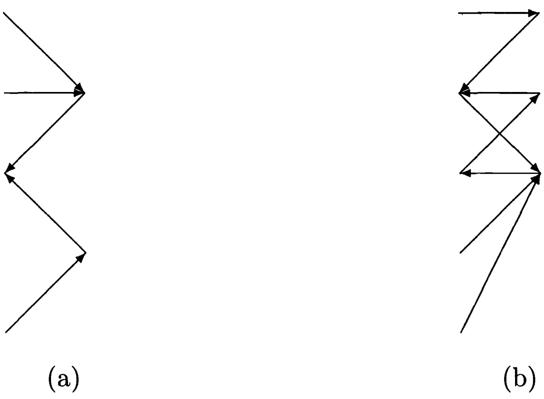


FIG. 2

at least 4. Adding to this function the generating function $T(x, y, z)$ of P-trees, we obtain $H(x, y, z)$, the generating function of all connected resource allocation graphs, with no isolated resources.

Lemma 2.3 then follows from Lemma 2.2 and a standard counting trick ([5, Lemma 3.2.16]) based on the decomposition of resource allocation graphs into components that shows that

$$G(x, y, z) = \exp(H(x, y, z)). \qquad \square$$

THEOREM 2.4. *We have*

$$a(m, n, q) = \binom{m-1}{q-n} m^n n^{q-n}.$$

Proof. If an acyclic resource allocation graph has m processes, n resources (none of which are free), and q edges, then it has $m+n-q$ components. The generating function for the set of acyclic resource allocation graphs with c components is $T(x, y, z)^c/c!$, and therefore

$$a(m, n, q) = \left[\frac{x^m y^n z^q}{m!n!}\right] \frac{T(x, y, z)^{m+n-q}}{(m+n-q)!}.$$

We evaluate this by using the multivariate Lagrange inversion theorem ([5, Thm. 1.2.9]). It is applied to the following formal power series in the ring $R[[x, y]]$, where $R = \mathbb{Z}[z]$:

$$\begin{array}{ll} f(x, y) = x^{m+n-q}, & \\ \phi_1(x, y) = e^{zy}, & \phi_2(x, y) = e^{zx}, \\ w_1(x, y) = T(x, y, z), & w_2(x, y) = U(x, y, z). \end{array}$$

By Lemma 2.1,

$$w_1 = x\phi_1(w_1, w_2) \quad \text{and} \quad w_2 = y\phi_2(w_1, w_2).$$

Therefore, by the multivariate Lagrange inversion theorem,

$$\begin{aligned} &[x^m y^n] f(w_1, w_2) \\ &= [x^m y^n] \left\{ f(x, y)\phi_1(x, y)^m \phi_2(x, y)^n \left\| \begin{array}{cc} 1 - \frac{x}{\phi_1(x,y)} \frac{\partial \phi_1(x,y)}{\partial x} & -\frac{y}{\phi_1(x,y)} \frac{\partial \phi_1(x,y)}{\partial y} \\ -\frac{x}{\phi_2(x,y)} \frac{\partial \phi_2(x,y)}{\partial x} & 1 - \frac{y}{\phi_2(x,y)} \frac{\partial \phi_2(x,y)}{\partial y} \end{array} \right\| \right\}. \end{aligned}$$

That is,

$$\begin{aligned}[x^m y^n]T(x,y,z)^{m+n-q} &= [x^m y^n]\{x^{m+n-q} e^{mzy} e^{nzx}(1-xyz^2)\}\\ &= \frac{m^n n^{q-n}}{n!(q-n)!}z^q - \frac{m^{n-1} n^{q-n-1}}{(n-1)!(q-n-1)!}z^q\\ &= \frac{m^{n-1} n^{q-n}(m+n-q)}{n!(q-n)!}z^q.\end{aligned}$$

Therefore

$$a(m,n,q) = \binom{m}{q-n} m^{n-1} n^{q-n}(m+n-q),$$

and the theorem follows. □

THEOREM 2.5. *We have*

$$g(m,n,q) = \binom{m}{q-n} m^n n^{q-n} \sum_i \binom{n}{i}\binom{q-n}{i} i!(-mn)^{-i}.$$

Proof. Since

$$g(m,n,q) = \left[\frac{x^m y^n z^q}{m!n!}\right] G(x,y,z),$$

by Lemma 2.3 we can use the multivariate Lagrange inversion theorem with

$$f(x,y) = (1-xyz^2)^{-1} e^{x-xyz^2}.$$

Proceeding as before, we have

$$\begin{aligned}[x^m y^n]G(x,y,z) &= [x^m y^n]\{e^{x-xyz^2} e^{mzy} e^{nzx}\}\\ &= [x^m y^n]\{e^{-xyz^2} e^{mzy} e^{(nz+1)x}\}\\ &= \sum_{i=0}^{\min(m,n)} \frac{(-1)^i m^{n-i}(nz+1)^{m-i}}{i!(n-i)!(m-i)!} z^{n+i}.\end{aligned}$$

Therefore

$$g(m,n,q) = \frac{m!n!}{(m+n-q)!} \sum_{i=0}^{\min(n,q-n)} \frac{(-1)^i m^{n-i} n^{q-n-i}}{i!(n-i)!(q-n-i)!},$$

and the theorem follows. □

The asymptotic behavior of $g(m,n,q)$ breaks down into three cases: n bounded, $q < 2n$ as $n \to \infty$, and $q \geq 2n$ as $n \to \infty$. Our methods are elementary. The survey by Bender [2] covers most of the ideas used here.

THEOREM 2.6. *If n is bounded as $m \to \infty$, then*

$$g(m,n,q) \sim \binom{m}{q-n} m^n n^{q-n} \left(1 - \frac{q}{mn}\right)^n.$$

Proof. By Theorem 2.5,

$$g(m,n,q) = \binom{m}{q-n} m^n n^{q-n} \sum_{i=0}^{\min(n,q-n)} \binom{n}{i} \frac{(q-n)!(-mn)^{-i}}{(q-n-i)!}. \tag{2.3}$$

Since n is bounded, for $i \leq n$ we have

$$\frac{(q-n)!}{(mn)^i(q-n-i)!} = \left(\frac{q}{mn}\right)^i + O\left(\frac{1}{m}\right).$$

If $q < 2n$, then the sum in (2.3) is asymptotic to 1, and the theorem follows. Thus let us assume $q \geq 2n$. Then

$$g(m,n,q) = \binom{m}{q-n} m^n n^{q-n} \left(1 - \frac{q}{mn}\right)^n \left(1 + O\left(\frac{1}{m}\right)\right),$$

and again the theorem follows. □

COROLLARY 2.7. *If $n \geq 2$ is bounded as $m \to \infty$, then*

$$p(m,n,q) = \left(\frac{n-1}{2n}\right)\left(\frac{q}{m}\right)^2 + O\left(\left(\frac{q}{m}\right)^3\right) + O\left(\frac{1}{m}\right).$$

Proof. From the proof of Theorem 2.6,

$$g(m,n,q) = \binom{m}{q-n} m^n n^{q-n} \left(1 - \frac{q}{mn}\right)^n \left(1 + O\left(\frac{1}{m}\right)\right).$$

Along with Theorem 2.4, this implies that

$$\frac{a(m,n,q)}{g(m,n,q)} = \left(1 - \frac{q}{m} + \frac{n}{m}\right)\left(1 - \frac{q}{mn}\right)^{-n}\left(1 + O\left(\frac{1}{m}\right)\right).$$

Now $q \leq m+n$, and for large enough m, $m+n \leq .9mn$, i.e., $q/(mn)$ is uniformly bounded away from 1. Therefore

$$\left(1 - \frac{q}{mn}\right)^{-n} = 1 + \frac{q}{m} + \frac{(n+1)q^2}{2m^2n} + O\left(\left(\frac{q}{m}\right)^3\right),$$

and we have

$$\begin{aligned} p(m,n,q) &= 1 - \left(1 - \frac{q}{m} + \frac{n}{m}\right)\left(1 + \frac{q}{m} + \left(\frac{n+1}{2n}\right)\left(\frac{q}{m}\right)^2 \right. \\ &\qquad \left. + O\left(\left(\frac{q}{m}\right)^3\right) + O\left(\frac{1}{m}\right)\right) \\ &= \left(\frac{n-1}{2n}\right)\left(\frac{q}{m}\right)^2 + O\left(\left(\frac{q}{m}\right)^3\right) + O\left(\frac{1}{m}\right). \end{aligned}$$ □

THEOREM 2.8. (a) *If $q < 2n$ as $n \to \infty$, then*

$$g(m,n,q) \sim \binom{m}{q-n} m^{2n-q}(mn-n)^{q-n}.$$

(b) *If* $q \geq 2n$ *as* $n \to \infty$, *then*

$$g(m,n,q) \sim \binom{m}{q-n} n^{q-2n}(mn+n-q)^n.$$

Proof. (a) Consider $q < 2n$ as $n \to \infty$. By Theorem 2.5,

$$g(m,n,q) = \binom{m}{q-n} m^{2n-q} \sum_i \binom{q-n}{i} \frac{(-1)^i (mn)^{q-n-i} n!}{(n-i)!}.$$

The absolute value of the ratio of consecutive terms in the sum is

$$\begin{aligned}\frac{(mn)^{q-n-i-1}}{(i+1)!(q-n-i-1)!(n-i-1)!} \Big/ \frac{(mn)^{q-n-i}}{i!(q-n-i)!(n-i)!} &= \frac{(q-n-i)(n-i)}{mn(i+1)} \\ &\leq \frac{q-n}{m(i+1)} \\ &\leq \frac{1}{i+1}.\end{aligned}$$

Thus the terms in the sum for $i \geq n^{1/3}$ are negligible. Using Stirling's formula to approximate the terms when $i < n^{1/3}$,

$$\begin{aligned}\frac{n!}{(n-i)!} &= \left(\frac{2\pi n}{2\pi(n-i)}\right)^{1/2} \frac{n^n e^{-n}}{(n-i)^{n-i} e^{i-n}} \left(1 + O\left(\frac{1}{n}\right)\right) \\ &= n^i \left(1 + \frac{i}{n-i}\right)^{n-i} e^{-i} \left(1 + O\left(\frac{1}{n}\right)\right) \\ &= n^i e^{-i^2/(2(n-i)) + O(i^3/(n-i)^2)} \left(1 + O\left(\frac{1}{n}\right)\right) \\ &= n^i \left(1 + O\left(\frac{1}{n^{1/3}}\right)\right).\end{aligned}$$

Again because the terms in the following sum are negligible for $i \geq n^{1/3}$,

$$\begin{aligned}g(m,n,q) &\sim \binom{m}{q-n} m^{2n-q} \sum_{i=0}^{q-n} \binom{q-n}{i} (mn)^{q-n-i}(-n)^i \\ &= \binom{m}{q-n} m^{2n-q}(mn-n)^{q-n}.\end{aligned}$$

(b) For $q \geq 2n$ we write

$$g(m,n,q) = \binom{m}{q-n} n^{q-2n} \sum_i \binom{n}{i} \frac{(-1)^i (mn)^{n-i}(q-n)!}{(q-n-i)!}.$$

As before,

$$\frac{(mn)^{n-i-1}}{(i+1)!(n-i-1)!(q-n-i-1)!} \Big/ \frac{(mn)^{n-i}}{i!(n-i)!(q-n-i)!} \leq \frac{1}{i+1},$$

and when $i < n^{1/3}$,

$$\begin{aligned}
\frac{(q-n)!}{(q-n-i)!} &= \left(\frac{2\pi(q-n)}{2\pi(q-n-i)}\right)^{1/2} \frac{(q-n)^{q-n}e^{n-q}}{(q-n-i)^{q-n-i}e^{n+i-q}} \left(1+O\left(\frac{1}{q-n}\right)\right) \\
&= (q-n)^i \left(1+\frac{i}{q-n-i}\right)^{q-n-i} e^{-i}\left(1+O\left(\frac{1}{n}\right)\right) \\
&= (q-n)^i e^{-i^2/(2(q-n-i))+O(i^3/(q-n-i)^2)} \left(1+O\left(\frac{1}{n}\right)\right) \\
&= (q-n)^i \left(1+O\left(\frac{1}{n^{1/3}}\right)\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
g(m,n,q) &\sim \binom{m}{q-n} n^{q-2n} \sum_{i=0}^{n} \binom{n}{i} (mn)^{n-i}(n-q)^i \\
&= \binom{m}{q-n} n^{q-2n}(mn+n-q)^n. \qquad \square
\end{aligned}$$

COROLLARY 2.9. *If $n \to \infty$, then*

$$p(m,n,q) = 1-\left(1-\frac{q-n}{m}\right)e^{(q-n)/m} + O\left(\frac{1}{m}\right) + O\left(\frac{1}{n^{1/3}}\right).$$

Proof. We assume $q < 2n$; the case when $q \geq 2n$ can be handled similarly. The proof of Theorem 2.8 shows that

$$g(m,n,q) = \binom{m}{q-n} m^n n^{q-n}\left(1-\frac{1}{m}\right)^{q-n}\left(1+O\left(\frac{1}{n^{1/3}}\right)\right),$$

and therefore

$$\begin{aligned}
\frac{a(m,n,q)}{g(m,n,q)} &= \left(1-\frac{q-n}{m}\right)\left(1-\frac{1}{m}\right)^{-(q-n)}\left(1+O\left(\frac{1}{n^{1/3}}\right)\right) \\
&= \left(1-\frac{q-n}{m}\right)e^{(q-n)/m+O((q-n)/m^2)}\left(1+O\left(\frac{1}{n^{1/3}}\right)\right) \\
&= \left(1-\frac{q-n}{m}\right)e^{(q-n)/m}\left(1+O\left(\frac{1}{m}\right)+O\left(\frac{1}{n^{1/3}}\right)\right),
\end{aligned}$$

since $q \leq m+n$. So

$$p(m,n,q) = 1-\left(1-\frac{q-n}{m}\right)e^{(q-n)/m} + O\left(\frac{1}{m}\right) + O\left(\frac{1}{n^{1/3}}\right). \qquad \square$$

COROLLARY 2.10. *If $(q-n)/m = o(1)$, then*

$$\lim_{m\to\infty} p(m,n,q) = 0.$$

If $\lim_{m+n\to\infty}(q-n)/m = \alpha$ for some $\alpha > 0$, then

$$\lim_{m+n\to\infty} p(m,n,q) = \beta$$

for some $\beta > 0$. In particular,

$$\beta = \begin{cases} \frac{n-1}{2n}\alpha^2 + O(\alpha^3) & \text{if } n \text{ is bounded,} \\ 1-(1-\alpha)e^{\alpha} & \text{if } n \to \infty. \end{cases}$$

3. Graphs with a specified number of processes and resources. We first study the case when all resources are allocated. As we would expect, there are many similarities to §2, and we do not give all the details of the proofs.

Let

$$\begin{aligned}
a(m,n) &= \text{the number of acyclic resource allocation graphs,}\\
g(m,n) &= \text{the number of resource allocation graphs,}\\
p(m,n) &= 1-\frac{a(m,n)}{g(m,n)}, \text{ the probability of deadlock for a randomly selected}\\
&\ \text{resource allocation graph,}\\
t(m,n) &= \text{the number of } P\text{-trees,}\\
u(m,n) &= \text{the number of } R\text{-trees, and}\\
c(m,n) &= \text{the number of bipartite cycles,}
\end{aligned}$$

where $|P| = m$ and $|R| = n$, and let $G(x,y)$, $T(x,y)$, $U(x,y)$, and $C(x,y)$ be the associated exponential generating functions.

LEMMA 3.1. *We have*

$$\begin{aligned}
T(x,y) &= xe^{U(x,y)},\\
U(x,y) &= ye^{T(x,y)},\\
C(x,y) &= \log\left((1-xy)^{-1}\right),\\
G(x,y) &= (1-T(x,y)U(x,y))^{-1}\exp\left(T(x,y)-T(x,y)U(x,y)\right).
\end{aligned}$$

THEOREM 3.2. $a(m,n) = m^n(n+1)^{m-1}$.

Proof. The proof is analogous to the proof of Theorem 2.4. We work in the ring $R[[x,y]]$, where $R = \mathbb{Z}$. Let

$$\begin{aligned}
&f(x,y) = e^x,\\
&\phi_1(x,y) = e^y, &&\phi_2(x,y) = e^x,\\
&w_1(x,y) = T(x,y), &&w_2(x,y) = U(x,y).
\end{aligned}$$

Applying the multivariate Lagrange inversion theorem,

$$\begin{aligned}
[x^m y^n]e^{T(x,y)} &= [x^m y^n]\left\{ e^x e^{my} e^{nx} \left\| \begin{matrix} 1-\frac{x}{e^y}\frac{\partial e^y}{\partial x} & -\frac{y}{e^y}\frac{\partial e^y}{\partial y} \\ -\frac{x}{e^x}\frac{\partial e^x}{\partial x} & 1-\frac{y}{e^x}\frac{\partial e^x}{\partial y} \end{matrix} \right\| \right\}\\
&= [x^m y^n]\{e^{my}e^{(n+1)x}(1-xy)\}\\
&= \frac{m^n(n+1)^m}{n!m!} - \frac{m^{n-1}(n+1)^{m-1}}{(n-1)!(m-1)!}\\
&= \frac{m^n(n+1)^{m-1}}{m!n!},
\end{aligned}$$

and the theorem follows. □

THEOREM 3.3. *We have*

$$g(m,n) = m^n(n+1)^m \sum_i \binom{m}{i}\binom{n}{i} i!(-m(n+1))^{-i}.$$

Proof. We use the multivariate Lagrange inversion theorem with

$$f(x,y)=(1-xy)^{-1}e^{x-xy}.$$

Proceeding as in Theorem 2.5, we have

$$\begin{aligned}[x^my^n]G(x,y)&=[x^my^n]\{e^{x-xy}e^{my}e^{nx}\}\\&=[x^my^n]\{e^{-xy}e^{my}e^{(n+1)x}\}\\&=\sum_{i=0}^{\min(m,n)}\frac{(-1)^im^{n-i}(n+1)^{m-i}}{i!(n-i)!(m-i)!},\end{aligned}$$

and the theorem follows. □

THEOREM 3.4. (a) *If m is bounded as $n\to\infty$, then*

$$g(m,n)\sim\left(\frac{m-1}{m}\right)^m m^n(n+1)^m.$$

(b) *If n is bounded as $m\to\infty$, then*

$$g(m,n)\sim\left(\frac{n}{n+1}\right)^n m^n(n+1)^m.$$

(c) *As $m,n\to\infty$,*

$$g(m,n)\sim e^{-1}m^n(n+1)^m.$$

Proof. We consider two cases: $m\leq n$ and $m>n$. Part (a) of the theorem is subsumed by the first case, while part (b) is covered by the second. Part (c) is included in both cases, but they give the same asymptotic formula when $m,n\to\infty$.

By Theorem 3.3, if $m\leq n$, then

$$g(m,n)=(n+1)^m m^{n-m}\sum_i\binom{m}{i}\frac{(-1)^im^{m-i}(n+1)^{-i}n!}{(n-i)!}.$$

For $i\leq n^{1/3}$, $n!/(n-i)!=n^i(1+O(n^{-1/3}))$, and larger terms are negligible. Therefore

$$\begin{aligned}g(m,n)&\sim(n+1)^m m^{n-m}\sum_{i=0}^m\binom{m}{i}(-1)^im^{m-i}\\&=(n+1)^m m^{n-m}(m-1)^m\\&\sim(n+1)^m m^n e^{-1}\quad\text{if } m\to\infty.\end{aligned}$$

The case $m>n$ is similar. □

COROLLARY 3.5. (a) *If m is bounded as $n\to\infty$, then*

$$p(m,n)=1-\left(\frac{m}{m-1}\right)^m\left(\frac{1+o(1)}{n}\right).$$

(b) *If n is bounded as $m\to\infty$, then*

$$p(m,n)=1-\left(\frac{n+1}{n}\right)^n\left(\frac{1+o(1)}{n+1}\right).$$

(c) *As* $m, n \to \infty$,

$$p(m,n) = 1 - e\left(\frac{1+o(1)}{n}\right).$$

These results imply that the probability of deadlock is high, even when $n \to \infty$. This may seem surprising, but an explanation is that most of the resource allocation graphs with m processes and n resources have a large number of edges, which makes deadlock very likely. As was shown in §2, the number of edges q plays a critical role in determining the probability of deadlock. In most actual systems, deadlock does not occur with great frequency. This may be due to a relatively low number of resource requests compared to the number of processes and resources. Thus sparse resource allocation graphs would seem to be of particular practical importance.

We conclude with asymptotic formulas for the case when free resources are allowed. Let $g^*(m,n)$ and $a^*(m,n)$ be the number of resource allocation graphs, respectively, acyclic resource allocation graphs, with m processes and n resources. Again there are different cases depending on the relative growth rates of m and n. First, however, we make a somewhat technical distinction between the cases $m \le n/(3\ln n)^6$ and $m > n/(3\ln n)^6$.

LEMMA 3.6. *If* $m \le n/(3\ln n)^6$ *as* $n \to \infty$, *then*

$$a^*(m,n) \sim \left(\frac{m}{m+1}\right)^{m-1} (m+1)^n (n+1)^{m-1}.$$

Proof. By Theorem 3.2, we must approximate

$$\sum_j \binom{n}{j} m^j (j+1)^{m-1}. \tag{3.1}$$

The ratio of consecutive terms is

$$\left(\frac{n-j}{j+1}\right) m \left(\frac{j+2}{j+1}\right)^{m-1}.$$

It is easy to see that the sequence of terms is unimodal, and if $j = mn/(m+1)$, the ratio is asymptotic to 1. Thus let us sum over a new index $k = j - mn/(m+1)$, so that $k = 0$ corresponds to the maximum term. Each term can then be written as

$$\begin{aligned} &m^{mn/(m+1)}\left(\frac{mn}{m+1}+1\right)^{m-1} \\ &\times \binom{n}{mn/(m+1)+k} m^k \left(\frac{mn/(m+1)+1+k}{mn/(m+1)+1}\right)^{m-1}. \end{aligned} \tag{3.2}$$

The first two factors do not depend on k, and

$$\left(\frac{mn}{m+1}+1\right)^{m-1} \sim \left(\frac{m}{m+1}\right)^{m-1} (n+1)^{m-1}.$$

We now examine the other factors. Using Stirling's formula,

$$
\begin{aligned}
&\binom{n}{mn/(m+1)+k} \\
&= \left(\frac{n}{2\pi(mn/(m+1)+k)(n/(m+1)-k)}\right)^{1/2} \\
&\quad\times \left(\frac{n}{mn/(m+1)+k}\right)^{mn/(m+1)+k} \left(\frac{n}{n/(m+1)-k}\right)^{n/(m+1)-k} \\
&\quad\times \left(1+O\left(\frac{1}{mn/(m+1)+k}+\frac{1}{n/(m+1)-k}\right)\right) \\
&\sim \frac{m+1}{(2\pi mn)^{1/2}} \left[\left(\frac{m+1}{m}\right)\left(\frac{n}{n+(m+1)k/m}\right)\right]^{mn/(m+1)+k} \\
&\quad\times \left[(m+1)\left(\frac{n}{n-(m+1)k}\right)\right]^{n/(m+1)-k} \qquad \text{if } k=o(n/m) \\
&= \frac{m+1}{(2\pi mn)^{1/2}}(m+1)^n m^{-mn/(m+1)-k} \\
&\quad\times \left[\frac{1}{1+(m+1)k/(mn)}\right]^{mn/(m+1)+k} \left[\frac{1}{1-(m+1)k/n}\right]^{n/(m+1)-k}.
\end{aligned}
$$

Now the first two factors above are independent of k, and $m^{-mn/(m+1)-k}$ cancels with other factors in (3.2). Approximating the bracketed terms above, we obtain

$$
\begin{aligned}
&\exp\left[\left(-\frac{(m+1)k}{mn}+\frac{1}{2}\left(\frac{(m+1)k}{mn}\right)^2+O\left(\left(\frac{(m+1)k}{mn}\right)^3\right)\right)(mn/(m+1)+k)\right] \\
&\quad\times \exp\left[\left(\frac{(m+1)k}{n}+\frac{1}{2}\left(\frac{(m+1)k}{n}\right)^2+O\left(\left(\frac{(m+1)k}{n}\right)^3\right)\right)(n/(m+1)-k)\right] \\
&= \exp\left[-k+\frac{(m+1)k^2}{2mn}+O\left(\frac{(m+1)^2k^3}{(mn)^2}\right)-\frac{(m+1)k^2}{mn}+\frac{(m+1)^2k^3}{2(mn)^2}\right. \\
&\qquad\left. +O\left(\frac{(m+1)^3k^4}{(mn)^3}\right)\right] \\
&\quad\times \exp\left[k+\frac{(m+1)k^2}{2n}+O\left(\frac{(m+1)^2k^3}{n^2}\right)-\frac{(m+1)k^2}{n}-\frac{(m+1)^2k^3}{2n^2}\right. \\
&\qquad\left. +O\left(\frac{(m+1)^3k^4}{n^3}\right)\right] \\
&\sim \exp\left[-\frac{(m+1)k^2}{2mn}-\frac{(m+1)k^2}{2}\right] \qquad \text{if } k=o((n/m)^{2/3}).
\end{aligned}
$$

Examining the last factor in (3.2), it equals

$$
\begin{aligned}
&\left(1+\frac{(m+1)k}{mn+m+1}\right)^{m-1} \\
&= \exp\left[\frac{(m-1)(m+1)k}{mn+m+1}-\frac{(m-1)(m+1)^2k^2}{2(mn+m+1)^2}+O\left(\frac{(m-1)(m+1)^3k^3}{(mn+m+1)^3}\right)\right] \\
&\sim 1 \quad \text{if } k=o(n/m).
\end{aligned}
$$

Altogether, we have shown that each term (3.2) with $k=o((n/m)^{2/3})$ is asymptotic to

$$\left(\frac{m}{m+1}\right)^{m-1} (m+1)^n(n+1)^{m-1}\frac{m+1}{(2\pi mn)^{1/2}}\exp\left[-\frac{(m+1)^2k^2}{2mn}\right].$$

Also, it is evident that each factor $\exp[-(m+1)^2k^2/(2mn)]$ with $k > (n/m)^{7/12}$ is $o(n^{-1})$. That is, the sum of the terms in

$$\sum_k \exp\left[-\frac{(m+1)^2k^2}{2mn}\right] \quad \text{for } k > (n/m)^{7/12}$$

is negligible. Thus the sum is asymptotic to

$$\frac{(2mn)^{1/2}}{m+1}\int_{-\infty}^{\infty} e^{-x^2}\,dx = \frac{(2\pi mn)^{1/2}}{m+1}$$

and

$$a^*(m,n) \sim \left(\frac{m}{m+1}\right)^{m-1} (m+1)^n(n+1)^{m-1}. \qquad \square$$

LEMMA 3.7. *If $n/(3\ln n)^6 < m = o(n)$ as $n \to \infty$, then*

$$a^*(m,n) \sim \left(m + e^{-(m-1)/(n+1)}\right)^n (n+1)^{m-1}.$$

Proof. Changing the index in (3.1) to $k = n - j$, the ratio of consecutive terms is

$$\left(\frac{n-k}{k+1}\right)\left(\frac{1}{m}\right)\left(\frac{n-k}{n-k+1}\right)^{m-1} \le \frac{n}{(k+1)m},$$

so the terms $k \ge m^{-2/3}n$ are negligible. Also,

$$\begin{aligned}(n-k+1)^{m-1} &= (n+1)^{m-1}\left(1-\frac{k}{n+1}\right)^{m-1} \\ &= (n+1)^{m-1}\exp\left[-\frac{(m-1)k}{n+1} - \frac{(m-1)k^2}{2(n+1)^2} + O\left(\frac{(m-1)k^3}{(n+1)^3}\right)\right] \\ &\sim (n+1)^{m-1}\exp\left[-\frac{(m-1)k}{n+1}\right] \quad \text{if } k = o(m^{-1/2}n).\end{aligned}$$

Therefore

$$\begin{aligned}a^*(m,n) &\sim (n+1)^{m-1}\sum_{k=0}^{n}\binom{n}{n-k} m^{n-k}\left(e^{-(m-1)/(n+1)}\right)^k \\ &= (n+1)^{m-1}\left(m + e^{-(m-1)/(n+1)}\right)^n. \qquad \square\end{aligned}$$

THEOREM 3.8. (a) *If $m = o(n)$ as $n \to \infty$, then*

$$a^*(m,n) \sim \left(\frac{m}{m+1}\right)^{m-1} (m+1)^n(n+1)^{m-1}.$$

(b) *If $m = \alpha n$ as $m, n \to \infty$, then*

$$a^*(m,n) \sim e^{e^{-\alpha}/\alpha} m^n (n+1)^{m-1}.$$

(c) *If $n = o(m)$ as $m \to \infty$, then*

$$a^*(m,n) \sim m^n (n+1)^{m-1}.$$

Proof. (a) If $m \le n/(3\ln n)^6$, then Lemma 3.6 applies immediately. If $n/(3\ln n)^6 < m = o(n)$, then we apply Lemma 3.7 and use the fact that

$$\begin{aligned}\left(m + e^{-(m-1)/(n+1)}\right)^n &\sim m^n \exp\left[\frac{ne^{-(m-1)/(n+1)}}{m}\right] \\ &\sim m^n \exp\left[\frac{n}{m} - 1\right] \\ &\sim (m+1)^n \left(\frac{m}{m+1}\right)^{m-1}.\end{aligned}$$

(b) is immediate from Lemma 3.7, and (c) follows from the fact that only the last term in (3.1) is significant. □

Similar proofs applied to the formulas for $g(m,n)$ in Theorem 3.4 give the following theorem.

THEOREM 3.9. (a) *If m is bounded as $n \to \infty$, then*

$$g^*(m,n) \sim \frac{m+1}{m}\left(\frac{m-1}{m+1}\right)^m (m+1)^n (n+1)^m.$$

(b) *If n is bounded as $m \to \infty$, then*

$$g^*(m,n) \sim \left(\frac{n}{n+1}\right)^n m^n (n+1)^m.$$

(c) *If $m = o(n)$ as $m, n \to \infty$, then*

$$g^*(m,n) \sim e^{-2}(m+1)^n (n+1)^m.$$

(d) *If $m = \alpha n$ as $m, n \to \infty$, then*

$$g^*(m,n) \sim e^{e^{-\alpha}/\alpha - 1} m^n (n+1)^m.$$

(e) *If $n = o(m)$ as $m, n \to \infty$, then*

$$g^*(m,n) \sim e^{-1} m^n (n+1)^m.$$

Obvious conclusions about the probability of deadlock can be derived. We give one statement that summarizes all possibilities.

COROLLARY 3.10. *We have*

$$1 - \frac{a^*(m,n)}{g^*(m,n)} = 1 - \frac{\beta(m,n)}{n},$$

where $\beta(m,n) \to e$ if $m, n \to \infty$.

4. Open problems.

1. What is the expected number of acyclic components in a random resource allocation graph? Since a process is not blocked if and only if it is the root of a P-tree, this is equivalent to the expected number of active processes.

2. Find realistic transition probabilities for resource allocation graphs, and calculate the expected time to deadlock. Is the ratio $(q-n)/m$ significant in determining the expected time until deadlock?

3. Find the expected size of the first cycle in a randomly evolving resource allocation graph, i.e., the expected number of processes involved in deadlock when it first occurs. This was recently solved for random graphs in [4].

4. Consider more general models of resource allocation, where processes can request several resources at a time, and there can be more than one instance of each resource.

5. Acknowledgment. The author thanks the Courant Institute of Mathematical Sciences, New York University, for its hospitality during a sabbatical visit when the final version of this article was prepared.

REFERENCES

[1] T. L. Austin, *The enumeration of point labelled chromatic graphs and trees*, Canad. J. Math., 12 (1960), pp. 535–545.

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

[3] P. Erdös and A. Rényi, *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 5 (1960), pp. 17–61.

[4] P. Flajolet, D. E. Knuth, and B. Pittel, *The first cycles in an evolving graph*, Discrete Math., 75 (1989), pp. 167–215.

[5] I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, John Wiley, New York, 1983.

[6] M. Hofri, *The deadlock problem in computing and communications systems—an annotated bibliography*, Tech. Report 500, Technion, Haifa, Israel, 1988.

[7] H. I. Scoins, *The number of trees with nodes of alternate parity*, Proc. Cambridge Philos. Soc., 58 (1962), pp. 12–16.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 474–483, August 1994

011

EVEN CYCLES IN DIRECTED GRAPHS*

F. R. K. CHUNG[†], WAYNE GODDARD[‡], AND DANIEL J. KLEITMAN[§]

Abstract. It is proved that every strongly connected directed graph with n nodes and at least $\lfloor (n+1)^2/4 \rfloor$ edges must contain an even cycle. This is best possible, and the structure of extremal graphs is discussed.

Key words. directed graphs, even cycles, permanents

AMS subject classifications. 05C, 15A

1. Introduction.

A directed graph G is a set of nodes $N(G)$ together with an edge set $E(G)$ consisting of ordered pairs of $N(G)$. A *path* in G from a node u to a node v is a sequence of distinct nodes $u = v_0, v_1, \ldots, v_t = v$, so that (v_i, v_{i+1}), $i = 0, \ldots, t-1$, are in $E(G)$. A path from u to v together with the edge (v, u) is called a *cycle*. If a cycle contains an even number of edges, it is said to be an even cycle. A *hamiltonian* cycle is a cycle that contains every node in the graph; a hamiltonian graph is one that has such a cycle. A graph G is said to be *strongly connected* if, for every pair of nodes u and v in $N(G)$, there is a path from u to v. In this paper, we consider directed graphs containing no loop (i.e., $(v,v) \notin E(G)$ for any v).

In this paper, we prove that every strongly connected graph on n nodes and at least $\lfloor (n+1)^2/4 \rfloor$ edges must contain an even cycle, thus settling a conjecture of Brualdi and Shader [4], [5]. This is best possible, and we give several examples of *edge-critical* graphs that are strongly connected directed graphs on n nodes and $\lfloor (n+1)^2/4 \rfloor - 1$ edges containing no even cycle. In particular, we characterize the edge-critical graphs that are hamiltonian.

Two simple examples of edge-critical graphs are the following.

1. The node set of H_n can be partitioned into three parts, A, B, and a node v, where $|A| = \lfloor (n-1)/2 \rfloor$ and $|B| = \lceil (n-1)/2 \rceil$. $E(H_n) = \{\, (a,b) : a \in A, b \in B \,\} \cup \{\, (v,a) : a \in A \,\} \cup \{\, (b,v) : b \in B \,\}$.

2. The node set of L_n consists of $v_0, v_1, \ldots, v_{n-1}$, which form a cycle. In addition, (v_i, v_j) is an edge if $i - j$ is a positive even number.

For a node v, the in-degree of v is the cardinality of $\{\, u : (u,v) \in E(G) \,\}$, and the out-degree of v is the cardinality of $\{\, u : (v,u) \in E(G) \,\}$. If for all v the in- and out-degree of v is d, we say that G is d-regular. We say nodes u and v are adjacent, or u is adjacent with v, or u is a neighbor of v, if either (u,v) or (v,u) is in $E(G)$. The adjacency matrix of a directed graph G on n nodes is the 0-1 $n \times n$ matrix M with $M(u,v) = 1$ if and only if $(u,v) \in E(G)$. For a subset S of $N(G)$, the induced subgraph of G on S has node set S and edge set $\{\, (a,b) \in E(G) : a, b \in S \,\}$. A node v is said to be a *cut point* if, by removing v and edges incident to v from G, the resulting graph is no longer strongly connected.

* Received by the editors January 23, 1992; accepted for publication (in revised form) May 26, 1993.

[†] Bellcore, 445 South Street, Morristown, New Jersey 07962.

[‡] Department of Mathematics, University of Pennsylvania, Philadelphia, Pennsylvania 19104. The work of this author was supported in part by Air Force Office of Scientific Research contract 89-0271.

[§] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was supported in part by Air Force Office of Scientific Research contract 89-0271.

Determining if a directed graph contains an even cycle is a surprisingly difficult problem that was first raised by Younger (see [21]). Klee, Ladner, and Manber [10] and Thomassen [14] showed that determining whether a prescribed edge is in a directed cycle of a particular parity is NP-complete. Nevertheless, the complexity of determining if a directed graph contains an even cycle remains unresolved. Surprisingly, there is a polynomial-time algorithm [10] to test whether a graph has a directed cycle of odd length. In [17] Thomassen gives a polynomial-time algorithm for deciding whether a planar directed graph contains an even cycle.

There are several results that give sufficient conditions for a directed graph to have an even cycle. Thomassen [14] proved that every directed graph on n nodes and minimum out-degree $\lfloor \log_2 n \rfloor + 1$ contains an even cycle. He also constructed graphs with minimum out-degree $\lfloor \frac{1}{2} \log_2 n \rfloor$ that do not have an even cycle. Alon and Linial [1] improved on this by showing that minimum out-degree $\log_2 n - \frac{1}{3} \log_2 \log_2 n + O(1)$ guarantees an even cycle. Alon and Linial [1] also proved that if the minimum out-degree is δ and the maximum in-degree is Δ, then a directed graph contains a directed cycle whose length is a multiple of k, provided that $(\Delta\delta + 1)(1 - 1/k)^{\delta} < \frac{1}{e}$. Friedland [8] showed that for $d \geq 7$, every d-regular directed graph contains an even cycle. In [16] Thomassen extended this to the case $d \geq 3$. There, Thomassen also showed that if for every pair of nodes u and v, there are at least three (node) disjoint paths joining u to v, then a directed graph contains an even cycle.

The problem of finding even cycles in a directed graph is closely related to the problem of converting some 1-entries of a (0,1)-matrix A to -1, so the resulting matrix B satisfies the property that the determinant of B is equal to the permanent of A (see [2]). Therefore, a hard problem [19] of computing the permanent is transformed into an easy one for some graphs. Little [11] and Seymour and Thomassen [13] gave characterizations of the matrices for which such conversion is possible. In particular, a directed graph G containing no even cycles satisfies

$$\text{determinant}(I + A) = \text{permanent}(I + A),$$

where A is the adjacency matrix of G, and I is the identity matrix.

The result in this paper implies that the largest number of 1's in an irreducible $n \times n$ (0,1)-matrix with the same value for its determinant and permanent (see [3], [4], [20]) is $\lfloor (n+1)^2/4 \rfloor + n - 1$.

2. Hamiltonian graphs.

Let G be a directed graph without an even cycle.

LEMMA 2.1. *Suppose there are three nodes a, b, and c, and (a,b), (b,c), and (a,c) are edges of G (so that (a,b,c) forms a "transitive" triple). Then all paths from c to a contain b, and thus b is a cut point.*

Proof. Suppose there is a path P from c to a that does not contain b. Then there must be an even cycle that is either adding (a,c) to P, or adding $(a,b),(b,c)$ to P. Therefore all paths from c to a must contain b. □

The following lemma establishes the result for hamiltonian graphs.

LEMMA 2.2. *Let T be a cycle in G. Then any path in T on m nodes spans at most $f(m) = \lfloor (m^2 + 2m - 3)/4 \rfloor$ edges.*

Proof. The proof is by induction on m. Let P denote a path $v, v_1, \ldots, v_{m-1}$ in cycle T. If v is adjacent with at most $\lceil m/2 \rceil$ nodes of P, then we are finished, since $f(m) - f(m-1) = \lceil m/2 \rceil$. We may therefore assume that v is adjacent with at least $\lceil m/2 \rceil + 1$ nodes of P. Then v must be connected to two consecutive nodes in P of the form v_{2i} and v_{2i+1}. The edge between v and v_{2i} must be (v_{2i}, v), and the edge between v and v_{2i+1} must be (v, v_{2i+1}), else there is an even cycle.

Observe that $P' = v, v_1, \ldots, v_{2i}$ and $P'' = v, v_{2i+1}, \ldots, v_{m-1}$ form paths and may be extended to cycles. Further, (v_{2i}, v, v_{2i+1}) is a transitive triple, so that the removal of v from G disconnects v_{2i} from v_{2i+1}. This means that all edges between the two paths P' and P'' not involving v must be directed from P' to P''. Now an edge (v_a, v_b) with $a < b$ requires a and b to have opposite parities, or it produces an even cycle. Thus there are at most $i(m - 2i - 1)$ such edges.

Since every edge in P is either of this form, or within P' or P'', P can have at most $f(2i+1) + f(m-2i) + i(m-2i-1) \leq f(m)$ edges. □

We now construct all edge-critical hamiltonian graphs on $2k+1$ nodes. Start with a directed triangle T_1 on nodes v_1, v_2, v_3. Then for $i = 2, \ldots, k$, create triangle T_i by introducing two new nodes, say v_{2i} and v_{2i+1}, and three edges, so that v_{2i} and v_{2i+1} form a directed triangle with some node a_{i-1} of triangle T_{i-1} (in that order). This yields a spine S_k on $2k+1$ nodes and $3k$ edges. Then, to construct G_k, take S_k and for every pair of distinct nodes $v \in T_i$ and $w \in T_j$, $i \leq j$, add (if necessary) the edge (v, w) if and only if the path from w to v in S_k has even length (number of edges).

It is easy to show by induction on k that G_k so constructed has $k(k+2)$ edges: except for a_{k-1}, every other node is adjacent with exactly one of v_{2k} and v_{2k+1}. We can also show that G_k is hamiltonian by induction on k. Let H denote the hamiltonian cycle of the subgraph on the first $k-1$ triangles, and x the predecessor of a_{k-1} in H. Then the path from a_{k-1} to x in the spine has even length; thus the same is true of the path from v_{2k} to x in S_k, and hence (x, v_{2k}) is an edge of G_k. Hence the edge (x, a_{k-1}) in H can be replaced by the three edges (x, v_{2k}), (v_{2k}, v_{2k+1}), and (v_{2k+1}, a_{k-1}), yielding the hamiltonian cycle of G_k.

It is less trivial to show that G_k has no even cycle. Observe that every a_i is a cut node: any path from T_{i+1} (and beyond) to T_i (and before) must pass through a_i. Let C be any cycle, and let v and w be nodes of C in the triangles of least and greatest index, respectively. Then the w-to-v portion of C is the path along S_k; say it has $p+1$ nodes. Let $v = h_0, h_1, \ldots, h_j = w$ denote the remainder of C. Let s_i be the distance from h_i to h_{i-1} in the spine $(i = 1, \ldots, j)$. It may be shown by induction on j that $\sum_i s_i = p + 3(j-1)$. However, all the s_i are even. Hence $p + j$ is odd, but this is the length of C.

We now show that every graph G on $2k+1$ nodes that is hamiltonian and edge critical has the above form. We do this by induction on k, so assume $k \geq 2$. G has maximum degree of at least $k+2$, so by the proof of the above lemma, G has a cut point v whose removal partitions the nodes into parts S_1 and S_2 such that all edges between the parts are directed from S_1 to S_2. Further, the proof shows that $S_1 \cup \{v\}$ and $S_2 \cup \{v\}$ induce edge-critical hamiltonian graphs G_1 and G_2. By the inductive hypothesis, G_1 and G_2 have the requisite form.

We must show that v is in the last triangle of G_1. This is immediate if G_1 is a triangle, so let the last two triangles of the spine S_m of G_1 be T_{m-1} and T_m. Let a denote the node that T_{m-1} and T_m have in common, y the out-neighbor of a in T_m, and x the in-neighbor of a in T_{m-1}. Then (x, a, y) forms a transitive triple in G_1 (the path from y to x in S_m has length four). Thus a must be on every path in G from y to x, in particular, on such a path that goes via S_2. Hence v must be in the last triangle of G_1. Similarly, v must be in the first triangle of G_2.

There can be an edge from $v \in S_1$ to $w \in S_2$ only if the path from w to v has even length. But to obtain the correct number of edges, we must then have all possible edges from S_1 to S_2. This means that G is one of the G_k constructed above.

3. The general result.

We prove the following theorem.

MAIN THEOREM. *A strongly connected directed graph G on n nodes without an even cycle contains at most $f(n) = \lfloor (n^2 + 2n - 3)/4 \rfloor$ edges.*

Further, such a G with $f(n)$ edges consists of a maximal hamiltonian edge-critical subgraph C (on $2r + 1$ nodes, say), and an acyclic complete bipartite graph on the remaining nodes, having parts as equal in size as possible, each node of which is connected to $r + 1$ of the nodes of C.

Roughly speaking, the proof of the main theorem can be described as follows. We partition the node set of our directed graph into a sequence of subsets, which we call *layers*, so that the subgraph induced by each initial segment of layers is strongly connected. Each layer consists of either the nodes of a directed cycle, the nodes of a (nontrivial) directed path, or a single node. Thus we refer to *cycle*, *path*, and *singleton* layers.

We establish first an upper bound on the number of edges inside a layer. We then establish an upper bound B on the number of edges between any two layers. Finally, we show that this bound B cannot be attained between every pair of layers; indeed, we derive an upper bound on the number of pairs of layers for which the bound B is attained.

3.1. Construction of layers. We assume throughout that the directed graph G has n nodes and no even cycle. The first layer L_0 consists of the nodes of a maximum cycle in G. Thereafter, we construct a sequence of layers L_j $(j \geq 1)$ of nodes of G as follows. Let S_j denote $\bigcup_{i \leq j} L_i$.

1. If there is a cycle T of $G - S_{j-1}$ such that there are edges between T and S_{j-1} in both directions, then we let the nodes of L_j be those of one such cycle of maximum length.

2. Otherwise, if there is a node v of $G - S_{j-1}$ such that there are edges between v and S_{j-1} in both directions, then we let L_j be $\{v\}$.

3. Otherwise, we let the nodes of L_j be those of a maximum length path $P = x \dots y$ in $G - S_{j-1}$, subject to there being an edge from S_{j-1} to x, and an edge from y to S_{j-1}.

Observe that L_j is (part of) a cycle in S_j and that S_j is strongly connected.

3.2. Excess. Now, we define the *excess* between two layers of cardinalities a and b as the number of edges between them minus $ab/2$. Similarly, we define the excess inside a layer of cardinality a as the number of edges within it, minus $\binom{a}{2}/2$.

To analyze the number of edges between layers, we consider an auxiliary undirected graph H whose nodes correspond to the layers in G. There is an edge between two nodes in H if and only if there is positive excess between the corresponding layers in G (i.e., the number of edges exceeds half the product of the layers' cardinalities). We establish a tight upper bound on the number of edges in H. In this regard, we define a forest F of the edges of H as follows: For each node w of H, we include in F the edge of H linking w to the node corresponding to the layer of smallest index, such that there are edges in both directions between the corresponding layers in G, if such a layer exists.

In order to prove the main theorem we prove the following three claims.

CLAIM 1. *The number of edges of G within any layer on m nodes is at most $f(m)$.*

CLAIM 2. *The excess in G between any two layers is at most 1. If equality holds, then there is an edge between the corresponding nodes in F. The excess between the first layer and a nonsingleton layer is at most -1.*

CLAIM 3. *The graph $H - F$ has no triangles, and the first layer is isolated in $H - F$.*

Actually, we have already established Claim 1 (see Lemma 2.2).

3.3. Proof of the main theorem. We show that the above three claims are sufficient to establish that the excess in G is at most $\frac{3n}{4} - \frac{3}{4}$. Let x denote the number of layers of odd cardinality, excluding L_0, and y the number of even cardinality. The contribution to the number of edges in G is in two parts.

(i) *Inside a layer.* In a layer of size m, there is an excess of at most $\frac{3m}{4} - \frac{3}{4}$ if m is odd, and at most $\frac{3m}{4} - 1$ if m is even, by Claim 1. Hence the excess inside layers is at most

$$3n/4 - y - 3x/4 - 3/4;$$

(ii) *Between layers.* We say that a potential edge e between two nodes of H is "odd" if the layers corresponding to both its ends have odd cardinality; otherwise it is "even." By Claim 2, if e is odd, then it contributes at most $+\frac{1}{2}$ to the excess in G if it is in H, and $-\frac{1}{2}$ otherwise. If e is even, then it contributes $+1$ if in F, and 0 otherwise. We may reassign the contributions and say: Any edge contributes $+1$ if it is in F; an odd edge contributes $+\frac{1}{2}$ if it is in $H - F$, and $-\frac{1}{2}$ otherwise. There are $\binom{x+1}{2}$ potential odd edges. By Claim 3, at most $x^2/4$ of these are in $H - F$. There are at most $x + y$ edges in F since it is a forest. Hence the excess between layers is at most

$$\frac{1}{2} \cdot \frac{x^2}{4} - \frac{1}{2}\left(\binom{x+1}{2} - \frac{x^2}{4}\right) + (x + y) = 3x/4 + y.$$

Thus the overall excess is at most $\frac{3n}{4} - \frac{3}{4}$.

We can also deduce the partial characterization of edge-critical graphs. The first layer L_0 is a cycle layer. By Claim 2, equality in the main theorem requires all layers except L_0 to be singleton layers. Also, a singleton layer must have positive excess with L_0; indeed, it must be joined to L_0 by an edge in F, which means that it has edges in both directions with L_0 in G. Further, $G - L_0$ must be complete bipartite having as equal as possible size parts (by Turán's theorem [18]). And by the construction of the layers, $G - L_0$ must be acyclic.

It remains to establish Claims 2 and 3.

4. Proof of Claim 2.

LEMMA 4.1. *Let T be any cycle $w_0, w_1, \dots, w_{2r}, w_0$ in G, and let v be a node in $G - T$. Then v can connect to only one pair of consecutive nodes in T, say (w_0, w_1). Thus there are at most $r + 1$ edges between v and T. If there are $r + 1$ edges, then v connects to w_0 and to $w_1, w_3, \dots, w_{2r-1}$.*

Proof. Let there be edges between v and both w_0 and w_1 in G. Of the four possible pairs of directions for these edges, one (viz., (w_0, v) and (v, w_1)) produces an even cycle with the rest of T.

Suppose that the two edges are (w_1, v) and (v, w_0), so that $\{v, w_0, w_1\}$ forms a directed cycle. It is immediate that, independent of direction, to avoid even cycles, all other edges between T and v involve only odd-indexed nodes of T.

Suppose now that both edges are directed from T toward v. Then by Lemma 2.1, removal of w_1 destroys all paths from v to w_0. Thus any path from v to w_1 is internally disjoint from T. Hence, all edges between v and T are directed toward v and, moreover, can exist only from odd-indexed nodes w_a in T. □

We define the *in-node* v of a layer L_j ($j \geq 1$) as follows: If L_j is a cycle layer, then v is any node such that there is an edge from S_{j-1} to v; if L_j is a path layer, then v is the first node on the path; and if L_j is a singleton layer, then v is the node itself. The *out-node* of L_j is defined analogously.

LEMMA 4.2. *Let node v in $G - S_j$ be adjacent with two consecutive nodes x and y in layer L_j.*

1. *If both edges are directed towards v, then L_j is a cycle layer, $j \geq 1$, y is the unique in-node of L_j, and any path from v to S_{j-1} avoids L_j.*

2. *If the edges go in opposite directions, then $\{v, x, y\}$ forms a directed triangle.*

Proof. If the edges go in opposite directions, then $\{v, x, y\}$ forms a directed triangle, otherwise there is an even cycle. Suppose both edges are directed towards v. Then any path P from v to x must go via y (by Lemma 2.1). Consider the first node w of P that is in S_j. If w is in L_j, then it must be y. But then we can replace the edge (x, y) by (x, v) and P, thereby contradicting the maximality of L_j. Thus w must be in S_{j-1}. Hence y must be the unique in-node from S_{j-1}, and so L_j must be a cycle layer. □

We say a node v of $G - S_j$ is *special* with respect to layer L_j if it is adjacent with more than half the nodes in L_j. Further, we say that v is *backtrack special* if it is in a directed triangle with consecutive nodes of L_j. It is *up special* if all the edges from L_j are directed towards v, and *down special* if they are directed away from v. The above lemma shows that there is no up- or down-special node with respect to the first layer L_0.

LEMMA 4.3. *Let L_j be a path layer with path $x \dots y$, and assume that node v of $G - S_j$ is special with respect to L_j, but not backtrack special. Then L_j has an odd number of nodes, and v is adjacent with every alternate node on L_j with (x, v) and (v, y) being edges. We say such a node v is detour special.*

Proof. By Lemma 4.2, if v is not backtrack special, then it has no consecutive neighbors on L_j. Thus L_j is odd, and v is adjacent with every alternate node starting with x.

Suppose edges between v and L_j go only one way, say toward v. Then consider a path P from v to S_{j-1}. If P is disjoint from L_j, then it contradicts the maximality of the path L_j, since edge (y, v) is in G. If it meets L_j after x, then again L_j can be lengthened. But if it meets x first, then there is a longer cycle—by using L_j, (y, v), and P—and this cycle is anchored (with y and x) to S_{j-1}. This is a contradiction.

Thus edges go both ways. Then the edges directed away from v come after those directed toward v, otherwise an even cycle results; thus it follows that (x, v) and (v, y) are edges. □

LEMMA 4.4. *If nodes v and w of $G - S_j$ are backtrack special with respect to L_j, then they do not lie together in a cycle outside S_j.*

Proof. Let nodes v and w be backtrack special with respect to L_j. We show that, for some parity π, there are paths in both directions between v and w that use only S_j, and whose lengths have parity π. This contradicts the nonexistence of even cycles in G if v and w are in a cycle in $G - S_j$.

Suppose L_j is a path layer. Then we may assume that the path is $\dots x_0 x_1 x_2 \dots$, that v is connected to x_0 and x_1, and that w is connected to x_i and x_{i+1} for $i \geq 0$. If $i = 0$, then there are odd paths between v and w in both directions through L_j; if $i = 1$, then there are even paths. Otherwise, suppose that i is even. Then v is connected to x_{i+1} (by Lemma 4.1). It is not possible to orient this edge without producing paths of the same parity in both directions between v and w through S_j. The case when i is odd, or when L_j is a cycle layer, is similar. □

LEMMA 4.5. *Let nodes v and w be consecutive in layer L_k and have a common neighbor x in lower layer L_j (i.e., $k > j$).*

1. *If both edges are directed away from x, then L_k is a cycle layer and v is the unique out-node of L_k.*

2. *If the edges go in opposite directions and L_k is a cycle layer, then $\{v, w, x\}$ forms a directed triangle.*

Proof. Suppose (x, v) and (x, w) are in G. Then any path P from w to x goes via v (by Lemma 2.1). Thus v is the unique out-node to S_{k-1}, and L_k must be a cycle layer. If the edges go in opposite directions and L_k is a cycle layer, then $\{v, w, x\}$ forms a directed triangle, otherwise an even cycle results. □

We define a (potential) edge between two nodes in H as *one way* or *two way*, depending on whether there are edges in one direction or in both directions between the corresponding layers in G.

We now prove Claim 2.

LEMMA 4.6. *Let layer L_k be above layer L_j (i.e., $k > j$). Then Table* 1 *gives upper bounds on the excess between L_j and L_k in G, depending on whether there are edges in G between the layers in one direction or in both directions.*

TABLE 1

L_k	L_j	One way	Two way	Comments
C	C	$+1/2$	$-1/2$	$-3/2$ excess, if $j = 0$
C	P	0	$+1$	0 excess in $H-F$
P	C	$+1/2$	0	-1 excess, if $j = 0$
P	P	0	0	
C	S	$+1/2$	$-1/2$	
S	C	$+1/2$	$+1/2$	+ve excess for $j = 0$ requires two way
S	P	0	$+1$	0 excess in $H-F$
P	S	$+1/2$	$+1/2$	
S	S	$+1/2$	–	

Proof. 1. *Cycle above cycle.* Suppose L_k has an up-special node u. Then by Lemma 4.2, all edges are directed towards L_k. Further, u must be adjacent with the unique in-node of L_j. Thus, there are at most $(|L_k|+1)/2$ up-specials (by Lemma 4.1), and the bound follows. Otherwise, the only chance of positive excess is to have a backtrack-special node, but there can be at most one such node by Lemma 4.4.

Suppose there is $-\frac{1}{2}$ excess and $j = 0$. Then L_k is a triangle $abca$ with c (say) backtrack special. Let cycle L_0 be $x_0, x_1, \ldots, x_{2r}, x_0$. Then a and b must have r neighbors on L_0. We claim that a cannot have two consecutive neighbors in L_0; for by Lemma 4.2, they would form a directed triangle with a, which yields a contradiction as in the proof of Lemma 4.4. Let c's consecutive neighbors in L_0 be x_0 and x_1. Then, by the lack of even cycles and the maximality of L_0, a cannot be adjacent with x_0 or x_2. Thus a must be adjacent with x_1. Similarly, b must be adjacent with x_0. These two edges cannot be in opposite directions (e.g., $x_1abx_0x_1$ would be a 4-cycle). So, without loss of generality, we may assume that (b, x_0) and (a, x_1) are edges. Thus (b, c, x_0) is a transitive triple, and all edges are directed from b to L_0. It then follows that (b, x_3) is an edge (indices modulo $2r + 1$). But that means there is a cycle of length $2r + 3$ in G, viz., $x_3x_4 \ldots x_1cabx_3$, which contradicts our choice for L_0.

2. *Cycle above path.* There is at most one backtrack-special node. Further, detour-special nodes, which are only possible if the path layer L_j has odd cardinality, are nonadjacent (by the lack of even cycles). Positive excess therefore requires edges to go both ways in G. By the layering strategy, L_k cannot be connected both ways to S_{j-1}; thus this edge is in F.

3. *Path above cycle.* Observe that by the layering strategy, no node in the path layer L_k can have edges both ways to L_j. Up-specials with respect to L_j are nonadjacent and the same with down-specials. Suppose L_k has a down-special node d and an up-special node u. By the maximality of L_j, u must come after d. By Lemma 4.2, the path from u to S_{j-1} must avoid L_j, and the path from S_{j-1} to d must avoid L_j, so that u and d lie in an odd cycle disjoint from L_j. However, then there is an even cycle, since there is a path of length two and one of length three from d to u using only L_j. Furthermore, if $j = 0$, then L_k can have no special node and so the excess is at most $-|L_k|/2$.

4. *Path above path.* By Lemma 4.3, special nodes have edges in both directions with L_j, which is impossible by the layering strategy.

5. *Cycle above singleton.* By Lemma 4.1, there is at most $+\frac{1}{2}$ excess. If there is positive excess and edges both ways between the singleton s and the cycle layer L_k, then s is in a directed triangle with L_k in G, which is impossible by the layering strategy.

6. *Singleton above cycle.* By Lemma 4.1, there is at most $+\frac{1}{2}$ excess. By Lemma 4.2, for $j = 0$, positive excess requires a two-way edge.

7. *Singleton above path.* By Lemma 4.3, if the node s in layer L_k is special, then it is connected both ways with L_j. By the layering strategy, s cannot be connected both ways to S_{j-1}; so this edge is in F.

8. *Path above singleton.* Suppose the node s in the singleton layer is adjacent with two consecutive nodes v and w in L_k. Node s cannot lie in a cycle outside S_{j-1}. Thus by Lemma 4.5, the only possibility is that the edges are (v, s) and (s, w), so that the removal of s disconnects v from w (by Lemma 2.1). However, then there must be a path either from w to s or from s to v outside S_{j-1}, so that s is in a cycle outside S_{j-1}, a contradiction. □

5. Proof of Claim 3.

LEMMA 5.1. *Assume all edges are directed from layer L_j to layer L_k in G, $j < k$, and there is positive excess. Let y be the in-node of L_j, and let v be the out-node of L_k.*

1. *Nodes v and y are unique.*

2. *It holds that $j \geq 1$. Any path from L_k to S_{j-1} avoids L_j. In particular, there is a path from v to S_{j-1} avoiding L_j and the rest of L_k.*

3. *Nodes v and y are adjacent. If L_j (L_k) is a cycle layer, then v (y) is adjacent with the predecessor of y (successor of v).*

Proof. 1. Node v is unique by definition, if L_k is a singleton or path layer. If L_k is a cycle layer, then it is unique by Lemma 4.5, and similarly for y.

2. Suppose there is a path from L_k to L_j in $G - S_{j-1}$. Then this path either contradicts the maximality of L_j, if that is a cycle layer, or it contradicts the fact that the node of L_j is not in a cycle outside S_{j-1}, if L_j is a singleton layer. There is a path from v to S_{j-1} avoiding the rest of L_k, since there is an edge from v to S_{k-1}.

3. This part follows from Lemmas 4.2 and 4.5. Node v is up special because more than half of the nodes of L_k are up special. All up-special nodes connect to y, and,

if L_j is a cycle layer, to its predecessor. A similar argument holds looking from L_j to L_k. □

LEMMA 5.2. *There is no triangle in H having all one-way edges.*

Proof. Suppose there is a triangle in H using one-way edges among nodes A, B, C with indices a, b, c. If it is a directed triangle, say $A \rightarrow B \rightarrow C \rightarrow A$ with c the maximum, then the B to A path through C contradicts the above avoidance property.

Assume then that they form a transitive triple (A, B, C) with $a < c$. Let v denote the in-node of A, and y the out-node of C. Then there is an edge from v to y by the above lemma. Further, there must be an even-length path from v to y through B (by part 3 of Lemma 5.1). Thus the removal of B must disconnect A from C.

However, this is impossible, because if $b > c$, then A and C lie together in the strongly connected subgraph S_c. If $a < b < c$, then by Lemma 5.1, there is a path from C to S_{b-1} (and hence A) that avoids B. Furthermore, if $b < a$, then by Lemma 5.1, any path from C to S_{b-1} avoids B, as do those from S_{b-1} to A. □

By Lemma 4.6, there are only two possibilities for two-way edges in $H-F$: odd path layer above singleton layer (PS), and singleton layer above cycle layer (SC).

LEMMA 5.3. *There is no triangle in $H-F$ with a* PS *two-way edge.*

Proof. Assume that path layer L_k (of odd cardinality) is above singleton layer L_j, and they are connected by a two-way edge in H. Then (by the proof of Lemma 4.6), the node s of L_j is connected to every alternate node in L_k (in particular, to the in-node x and out-node y of L_k). Since the edges directed towards s must come before those directed away from s (by the lack of even cycles), (x, s) and (s, y) must be edges in G. Further, any path from y to S_{j-1}, or from S_{j-1} to x, must avoid s, since s is not in a cycle outside S_{j-1}.

Suppose L_j and L_k are in a triangle with layer L in $H-F$. If L is also a singleton layer, with, say, node t, and connected to L_k by a two-way edge in H, then by the above observation, L_k must lie in an odd cycle disjoint from s and t. However, x and y are also connected by a path of length three through s and t, which yields an even cycle.

Hence we may assume that all edges between L and L_k go the same way, say toward L_k. Then there is a path from y directly to the subgraph S below all three layers. Also, because there is a path from L to s in $G - S$, by Lemma 5.1, all edges are directed from L to s. This yields a transitive triple (v, s, y) where v is the in-node of L. By Lemma 5.1 (or otherwise), there is a path direct from S to v. This yields an even cycle. □

LEMMA 5.4. *There is no triangle in $H-F$ with an* SC *two-way edge.*

Proof. Assume singleton layer L_k, with node s, and cycle layer L_j $(k > j)$, are joined in $H-F$ by a two-way edge, and that this edge is in a triangle with layer L in $H-F$. Then we claim that all edges between L and $L_j \cup L_k$ go the same way. For if L_k is joined to L by a two-way (SC) edge in H, then L must be one way with L_j, and then either the L_j to L, or the L to L_j path through s in G contradicts Lemma 5.1. A similar contradiction results if L is two way with L_j, or if L_k and L_j are both one way with L but in different directions. So we may assume that all edges are directed towards L.

The edge of F incident with L_k connects to a layer in the subgraph S below all three layers. (By definition, the edge goes below L_j; by Lemma 5.1, it must go below L.) A path also exists direct from the out-node y of L to S, and an edge exists from s to y. This means that (s, y) lies in an odd cycle disjoint from L_j. But there is also an even-length path from s to y using only L_j (by part 3 of Lemma 5.1). This yields a contraction. □

We have established Claim 3 and hence the proof of the main theorem is complete.

6. Open problems. There are many unsolved problems about even cycles in directed graphs, two of which we mention here. The most obvious one is the problem of deciding if a directed graph contains an even cycle. In [17] Thomassen gives a polynomial algorithm for deciding whether a planar directed graph contains an even cycle.

Erdös and Pósa [6], [7] proved that every *undirected* graph contains either k disjoint cycles or contains $ck \log k$ nodes that must meet all cycles, for some constant c. Recently, McCuaig [12] proved a conjecture of Gallai [9] by showing that every directed graph contains either two disjoint cycles or three nodes meeting all cycles. Does there exist [20] a number $f(k)$ for every integer k such that a directed graph contains either k disjoint cycles or a set of $f(k)$ nodes meeting all cycles?

Acknowledgments. The authors thank Zoltan Furedi, John Goldwasser, Hong-Jia Lai, Paul Seymour, and C.Q. Zhang for very helpful discussions.

REFERENCES

[1] N. Alon and N. Linial, *Cycles of length* 0 *modulo k in directed graphs*, J. Combin. Theory Ser. B, 47 (1989), pp. 114–119.

[2] R. Brualdi, *Counting permutations with restricted positions: Permanents of* (0, 1) *matrices*, Linear Algebra Appl., 104 (1988), pp. 173–183.

[3] ———, *Graphs and matrices and graphs*, Congr. Numer., 83 (1991), pp. 129–145.

[4] R. Brualdi and B. Shader, *Cutsets in bipartite graphs.* Linear and Multilinear Algebra, to appear.

[5] ———, *On sign-nonsingular matrices and the conversion of the permanent into the determinant*, in Applied Geometry and Discrete Mathematics, American Mathematical Society, Providence, RI, 1991, pp. 117–134.

[6] P. Erdös and L. Pósa, *On the maximal number of disjoint circuits in a graph*, Publ. Math. Debrecen, 9 (1962), pp. 3–12.

[7] ———, *On independent circuits contained in a graph*, Canad. J. Math., 17 (1965), pp. 347–352.

[8] S. Friedland, *Every 7-regular digraph contains an even cycle*, J. Combin. Theory Ser. B, 46 (1989), pp. 249–252.

[9] T. Gallai, *Problem* 6, in Theory of Graphs (Proc. Colloq., Tihany, 1966), Academic Press, New York, 1968, p. 362.

[10] V. Klee, R. Ladner, and R. Manber, *Signsolvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.

[11] C. Little, *A characterization of convertible* (0, 1)*-matrices*, J. Combin. Theory Ser. B, 18 (1975), pp. 187–208.

[12] W. McCuaig, *Intercyclic digraphs*, in Contemporary Mathematics, Vol. 147, American Mathematical Society, Providence, RI, 1993, pp. 203–245.

[13] P. Seymour and C. Thomassen, *Characterization of even directed graphs*, J. Combin. Theory Ser. B, 42 (1987), pp. 36–45.

[14] C. Thomassen, *Even cycles in directed graphs*, European J. Combin., 6 (1985), pp. 85–89.

[15] ———, *Sign-nonsingular matrices and even cycles in directed graphs*, Linear Algebra Appl., 75 (1986), pp. 27–41.

[16] ———, *The even cycle problem for directed graphs*, J. Amer. Math. Soc., 5 (1992), pp. 217–229.

[17] ———, *The even cycle problem for planar digraphs*, J. Algorithms, 15 (1993), pp. 61–75.

[18] P. Turán, *Egy gráfelméletei szélsöékfeladatról*, Matem. Physikai Lapok, 48 (1941), pp. 436–452.

[19] L. Valiant, *The complexity of computing the permanent*, Theoret. Comput. Sci., 8 (1979), pp. 189–201.

[20] V. Vazirani and M. Yannakakis, *Pfaffian orientations,* 0-1 *permanents and even cycles in directed graphs*, Discrete Appl. Math., 25 (1989), pp. 179–190.

[21] D. Younger, *Graphs with interlinked directed circuits*, in Proc. Midwest Sympos. on Circuit Theory, Vol. 2, 1973, pp. XVI2.1–XVI2.7.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 484–492, August 1994

ON THE SIZE OF WEIGHTS FOR THRESHOLD GATES*

JOHAN HÅSTAD†

Abstract. It is proved that if n is a power of 2, then there is a threshold function on n inputs that requires weights of size around $2^{(n \log n)/2-n}$. This almost matches the known upper bounds.

Key words. threshold function, computational complexity, neural nets

AMS subject classification. 68R05

1. Introduction.

One very interesting computational element is that of a threshold gate. A threshold gate of n inputs is specified by a set of *weights* $w_1, w_2, \ldots, w_n$ and a *threshold* t. On input $x_1, x_2, \ldots, x_n$ it outputs $\mathrm{sign}(\sum_{i=1}^{n} w_i x_i - t)$. In this notation we assume that the gate computes a function $\{-1,1\}^n \mapsto \{-1,1\}$, but the set $\{-1,1\}$ could be replaced by any two element set (e.g., by $\{0,1\}$).

Threshold circuits, i.e., circuits that contain threshold gates, have been studied extensively. On the more theoretical side, many upper and lower bounds on the power of small depth threshold circuits have been established. For a discussion of these results we refer to [1] and its references. It is striking to note that there are no known strong lower bounds for general depth-2 threshold circuits. On the more applied side we note that threshold circuits have many similarities with neural networks. We refer to [2], [5], [6] for more information on these connections and the area in general.

For both types of investigation mentioned above it is important to understand what conditions can be put on the weights w_i. Since some finite amount of precision is always sufficient, it is easy to see that we can assume that the weights are integers. Furthermore, it has been proved many times (one early source is [4]) that if we have a function with n inputs, then $|w_i| \leq 2^{-n}(n+1)^{(n+1)/2}$ is sufficient. Since there are at least $2^{n^2/2}$ different threshold functions [4], [8], [9], there are some functions that require $\max |w_i|$ on the order of $2^{n/2}$. On the other hand, there are at most 2^{n^2} threshold functions [3], [7], and hence these are essentially the best bounds that can be proved by this type of simple counting.

There are also known explicit functions that require weights of size at least c^n for some constant $c > 1$. In particular, if we let the input encode two numbers in binary and ask which number is greater, then a lower bound of essentially $2^{n/2}$ holds. There are other, slightly more complicated, explicit functions giving a value of c up to $(1+\sqrt{5})/2$ [6].

The above-mentioned bounds imply that $\Omega(n)$ bits are sometimes needed and that $O(n \log n)$ bits are always sufficient to specify the individual weights. The gap between these two bounds are not substantial enough to matter greatly in arguments of complexity theory, because in general we are only interested in whether the quantities are polynomial in size. However the gap is rather large, and the goal of this paper is to bring the two bounds closer together. We do this by improving the lower bound for an explicit function F_n. We prove that when $n \geq 8$ is a power of 2, this function requires $|w_i| \geq (1/2n)e^{-4n^{\beta}} 2^{(n \log n)/2-n}$ for all i, where $\beta = \log_2 \frac{3}{2}$. Comparing this to the known upper bounds, we see that it is essentially tight.

* Received by the editors August 17, 1992; accepted for publication (in revised form) June 4, 1993.

† Department of Numerical Analysis and Computing Science, Royal Institute of Technology, 100 44 Stockholm, Sweden.

The outline of the paper is as follows. In §2 we define our function and prove the lower bound on the size of the weights needed to realize this function. In §3 we recall the proof for the upper bound on the weights, and we end with some final comments in §4.

2. A function requiring large weights. Let us assume that n is a power of two and that $n = 2^m$. We use $\{-1, 1\}$ notation throughout and we think of vectors in $\{-1, 1\}^n$ as functions from $\{-1, 1\}^m$ to $\{-1, 1\}$. This convention makes us use two types of functions: those on m variables and those on n variables. We use the former as inputs to the latter. To decrease the possibility of confusion we reserve capital letters for functions on n inputs.

For $\alpha \subseteq [m] = \{1, 2, \ldots, m\}$, let φ_α be the character function, i.e., $\varphi_\alpha(x) = \prod_{i \in \alpha} x_i$, where we let $\varphi_\emptyset$ be the function that is identically 1. Choose an ordering of $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ of the sets such that the following conditions hold:

1. $|\alpha_i| \leq |\alpha_j|$ for $i \leq j$;
2. $|\alpha_i \Delta \alpha_{i+1}| \leq 2$ for all i, where $|\alpha_i \Delta \alpha_{i+1}|$ is the symmetric difference of the two sets α_i and α_{i+1}.

This implies that $\alpha_0 = \emptyset$ and $\alpha_1, \alpha_2, \ldots, \alpha_m$ are the singletons, and that $|\alpha_i \Delta \alpha_{i+1}| = 1$ when $|\alpha_{i+1}| = |\alpha_i| + 1$, while $|\alpha_i \Delta \alpha_{i+1}| = 2$ otherwise.

LEMMA 2.1. *There is an ordering that satisfies conditions* 1 *and* 2 *above.*

Proof. Assume that we have an ordering of all sets containing at most d elements satisfying conditions 1 and 2, and an ordering of the set with $d + 1$ elements that satisfies condition 2. If we think of these orderings as lists, we can concatenate them. This might create an illegal ordering in that condition 2 might not be satisfied when $|\alpha_i| = d$ and $|\alpha_{i+1}| = d+1$. However, if we simply permute the names of the elements in the sets of size $d + 1$ we can take care of this condition and the concatenated list will satisfy both conditions.

The above reasoning implies that the only problem is to find an ordering of the d-element subsets of $[m]$ for any $d \leq m$. We prove this by an induction over d and m. The base cases $d = 1$ and $d = m$ are obvious. In the general case we first list all the sets of size d containing m, and then all the other sets. This first part of the list is essentially all $(d - 1)$-element subsets from $[m - 1]$, while the second part consists of all d-element subsets of the same set. Both can be given appropriate orderings by induction, and thus the only problem is the connection between the two sublists. However, as above, by permuting the names of the elements in the second list, this connection can be made to satisfy condition 2. This completes the induction step and hence the lemma follows. □

Let (f, g) denote the inner product of the functions f and g. Define $F(f) : \{-1, 1\}^n \mapsto \{-1, 1\}$ as $\operatorname{sign}((f, \varphi_{\alpha_i}))$, where i is the largest index such that $(f, \varphi_{\alpha_i}) \neq 0$. This function is a threshold function since

$$F(f) = \operatorname{sign}\left(\sum_{i=0}^{n-1} (n+1)^i (f, \varphi_{\alpha_i})\right)$$

and (f, φ_{α_i}) is a linear function in the values $f(j)$. This expression is correct since $|(f, \varphi_{\alpha_i})| \leq n$ for all i.

We want to prove that if

$$F(f) = \operatorname{sign}\left(\sum_{j=0}^{n-1} w_j f(j) - t\right),$$

then one of w_j is large. First let us observe the following simple lemma.

LEMMA 2.2. *We can assume that $t = 0$.*

Proof. Note that $F(f) = -F(-f)$. Hence $|\sum_{j=0}^{n-1} w_j f(j)| > |t|$ for any f, and we can set $t = 0$ without changing the function. □

It will be easier to work with expressions of the form

$$\operatorname{sign}\left(\sum_{i=0}^{n-1} w_i'(f, \varphi_{\alpha_i})\right).$$

This can be done with the following lemma.

LEMMA 2.3. *We have*

$$\sum_{j=0}^{n-1} w_j f(j) = \sum_{i=0}^{n-1} w_i'(f, \varphi_{\alpha_i})$$

for all f if and only if

$$w_i' = \frac{1}{n}\sum_{j=0}^{n-1} w_j \varphi_{\alpha_i}(j)$$

and

$$w_j = \sum_{i=0}^{n-1} w_i' \varphi_{\alpha_i}(j).$$

Proof. The second statement follows from rearranging the terms. To see the first, note that by Fourier inversion,

$$f(j) = \frac{1}{n}\sum_{i=0}^{n-1} \varphi_{\alpha_i}(j)(f, \varphi_{\alpha_i}).$$

This implies

$$\sum_{j=0}^{n-1} w_j f(j) = \sum_{j=0}^{n-1} w_j \frac{1}{n}\sum_{i=0}^{n-1} \varphi_{\alpha_i}(j)(f, \varphi_{\alpha_i}) = \frac{1}{n}\sum_{i=0}^{n-1}(f, \varphi_{\alpha_i})\sum_{j=0}^{n-1} w_j \varphi_{\alpha_i}(j),$$

and the lemma follows. □

We will establish that F requires large weights, and in particular we have the following theorem.

THEOREM 2.4. *Assume that n is a power of 2 and*

$$F(f) = \operatorname{sign}\left(\sum_{i=0}^{n-1} w_i(f, \varphi_{\alpha_i})\right),$$

where w_i are integers. Then $w_{n-1} \geq e^{-4n^\beta} 2^{(n \log n)/2-n}$, where $\beta = \log(\frac{3}{2})$.

Before we prove Theorem 2.4, let us first deduce the corresponding result when using the normal representation of threshold gates.

THEOREM 2.5. *Assume that n is a power of 2 and*

$$F(f) = \operatorname{sign}\left(\sum_{j=0}^{n-1} w_j f(j)\right),$$

where w_j are integers. Then for some j, we have $|w_j| \geq \frac{1}{n}e^{-4n^\beta}2^{(n\log n)/2-n}$, where $\beta = \log(\frac{3}{2})$.

Proof of Theorem 2.5. If we have an expansion of the kind given in the theorem, then by Lemma 2.3, setting

$$w_i' = \frac{1}{n}\sum_{j=0}^{n-1} w_j \varphi_{\alpha_i}(j),$$

we convert it to an expansion of the form considered in Theorem 2.4. The w_i' may not be integers, but nw_i' are integers for all i. Multiplying every weight by the same integer does not change the function and hence by Theorem 2.4, $nw_{n-1}' \geq e^{-4n^\beta}2^{(n\log n)/2-n}$. This is equivalent to saying that

$$\sum_{j=0}^{n-1} w_j \varphi_{\alpha_{n-1}}(j) \geq e^{-4n^\beta}2^{(n\log n)/2-n},$$

and the theorem follows. □

Let us now prove Theorem 2.4.

Proof of Theorem 2.4. Let us start by an easy observation.

LEMMA 2.6. *For any i, $w_i > 0$.*

Proof. This follows from setting $f = \varphi_{\alpha_i}$ and noting that $F(f) = 1$ for such f. □

By choosing a suitable sequence of test functions f we prove that w_i must grow exponentially. Since the ordering of the α_i is not explicit, we sometimes use the notation w_α, which should be read as w_i where i is chosen such that $\alpha_i = \alpha$.

LEMMA 2.7. *Suppose $|\alpha_{i+1}| = |\alpha_i| = k$ where $2 \leq k \leq n-1$, and that $\alpha_i \Delta \alpha_{i+1} = \{a, b\}$. Let $v \in \{-1,1\}^m$ be any point with $v_a = v_b$. Then*

$$w_{i+1} > (2^{k-1}-1)w_i - \sum \varphi_{\alpha_i}(v)\varphi_\alpha(v)w_\alpha,$$

where the sum extends over all α such that $\alpha \subset \alpha_i \bigcup \alpha_{i+1}$, and α contains exactly one of a and b and α is equal to neither α_i nor α_{i+1}.

Proof. Let us assume that $\alpha_i = \{1,2,\ldots,k\}$ and $\alpha_{i+1} = \{1,2,\ldots,k-1,k+1\}$. Let v^1 be a vector of length $k+1$ that satisfies $v_j^1 = v_j$ for $1 \leq j \leq k+1$. Furthermore, let v^2 be a similar vector such that $v_j^2 = v_j$ for $j < k$, and $v_j^2 = -v_j$ for $j = k$ and $j = k+1$. Define the following function on the first $k+1$ variables:

$$f(w) = \begin{cases} \varphi_{\alpha_i}(w) & \text{if } w = v^1 \text{ or } w = v^2, \\ -\varphi_{\alpha_i}(w) & \text{otherwise.} \end{cases}$$

We extend f to a function of m variables by ignoring the rest of the variables. First note that $(f, \varphi_\alpha) = 0$ for any α that contains an element larger than $k+1$. We have that $(f, \varphi_{\alpha_i}) = 2^{m-k-1}(4 - 2^{k+1})$, while for other $\alpha \subseteq \{1,2,\ldots,k+1\}$ we have

$$\begin{aligned}(f, \varphi_\alpha) &= (-\varphi_{\alpha_i}, \varphi_\alpha) + 2^{m-k}\varphi_{\alpha_i}(v^1)\varphi_\alpha(v^1) + 2^{m-k}\varphi_{\alpha_i}(v^2)\varphi_\alpha(v^2) \\ &= 2^{m-k}\varphi_{\alpha_i}(v^1)\varphi_\alpha(v^1)(1 + \varphi_{\alpha_i}(v^1v^2)\varphi_\alpha(v^1v^2)),\end{aligned}$$

where v^1v^2 is the pointwise product of v^1 and v^2. Since this vector is 1 except for coordinates k and $k+1$, we obtain a nonzero inner product if and only if $\alpha \subset$

$\{1, 2, \ldots, k+1\}$ and α contains exactly one of the elements k and $k+1$. The only two sets of size k with these properties are α_{i+1} and α_i, while all other sets with these properties have cardinality at most $k-1$. Note that by property 1 of our ordering, all these sets appear before α_i, and this implies that $F(f) = \text{sign}((\varphi_{\alpha_{i+1}}, f)) = 1$. Writing this statement as an inequality of the weights yields the inequality of the lemma. □

Next we have the following lemma.

LEMMA 2.8. *Suppose* $|\alpha_{i+1}| = 1 + |\alpha_i| = k$, *where* $2 \leq k \leq n-1$ *and* $\alpha_{i+1} = \alpha_i \bigcup \{a\}$. *Then, for any vector* v *with* $v_a = 1$, *we have*

$$w_{i+1} > (2^{k-1} - 1)w_i - \sum_{\alpha \subset \alpha_{i+1}, \alpha = \alpha_i} \varphi_{\alpha_i}(v)\varphi_\alpha(v)w_\alpha.$$

Proof. Let us assume that $\alpha_i = \{1, 2, \ldots, k-1\}$ and $\alpha_{i+1} = \{1, 2, \ldots, k\}$, and let v^1 be the vector of length k that satisfies $v^1_j = v_j$ for $1 \leq j \leq k$. Define the following function on the first k variables:

$$f(w) = \begin{cases} \varphi_{\alpha_i}(w) & \text{if } w = v^1, \\ -\varphi_{\alpha_i}(w) & \text{otherwise.} \end{cases}$$

Extend f to a function of m variables by ignoring the rest of the variables. Again $(f, \varphi_\alpha) = 0$ for any α that contains an element larger than k. Clearly $(f, \varphi_{\alpha_i}) = 2^{m-k}(2 - 2^k)$, while for other $\alpha \subset \{1, 2, \ldots, k\}$ we have

$$(f, \varphi_\alpha) = (-\varphi_{\alpha_i}, \varphi_\alpha) + 2^{m+1-k}\varphi_{\alpha_i}(v^1)\varphi_\alpha(v^1) = 2^{m+1-k}\varphi_{\alpha_i}(v^1)\varphi_\alpha(v^1).$$

Again $F(f) = \text{sign}((\varphi_{\alpha_{i+1}}, f)) = 1$, and writing out the corresponding inequality for the weights gives the lemma. □

Using the above two lemmas we establish the main lemma for the proof of Theorem 2.4.

LEMMA 2.9. *For each* i *such that* $|\alpha_{i+1}| \geq 2$, *we have* $w_{i+1} > (2^{|\alpha_{i+1}|-1} - 1)w_i$. *Furthermore, if* $\alpha = \{a, b\}$, *then*

$$w_\alpha > w_{\{a\}} + w_{\{b\}} + w_0.$$

Proof. We establish the lemma by induction over i, and we must handle the cases with $|\alpha_{i+1}|$ small separately.

Let us first establish the lower bounds for w_α when $|\alpha| = 2$. Suppose $a = 1$ and $b = 2$, and construct the function of the proof of Lemma 2.8 when $\alpha_{i+1} = \alpha$, $\alpha_i = \{1\}$, and $v^1 = (-1, 1)$. This shows that

$$w_\alpha > w_{\{1\}} + w_{\{2\}} + w_0.$$

Let us now establish the lemma when $|\alpha_{i+1}| = |\alpha_i| = 2$. Suppose $\alpha_{i+1} = \{1, 3\}$ and $\alpha_i = \{1, 2\}$. Then apply Lemma 2.7 with $v^1 = (-1, 1, 1)$. This gives

$$w_{i+1} > w_i + w_{\{2\}} + w_{\{3\}}.$$

Next suppose $|\alpha_{i+1}| = 3$ and $|\alpha_i| = 2$. We might assume that $\alpha_{i+1} = \{1, 2, 3\}$ and $\alpha_i = \{1, 2\}$. Apply Lemma 2.8 with $v^1 = (-1, -1, 1)$. This gives

$$w_{i+1} > 3w_i + w_{\{1,3\}} + w_{\{2,3\}} + w_{\{1\}} + w_{\{2\}} - w_{\{3\}} - w_0 > 3w_i,$$

where the last inequality follows from the already established bounds.

Let us now take the case $|\alpha_{i+1}| = |\alpha_i| = 3$, where we assume that $\alpha_{i+1} = \{1, 2, 4\}$ and $\alpha_i = \{1, 2, 3\}$. Apply Lemma 2.7 with $v^1 = (-1, -1, 1, 1)$. This gives

$$w_{i+1} > 3w_i + w_{\{1,3\}} + w_{\{2,3\}} + w_{\{1,4\}} + w_{\{2,4\}} - w_{\{3\}} - w_{\{4\}} > 3w_i.$$

Now consider the case when $|\alpha_{i+1}| = k$ and $|\alpha_i| = k - 1$, where $k \geq 4$. We can assume that $\alpha_{i+1} = \{1, 2, \ldots, k\}$ and $\alpha_i = \{1, \ldots, k-1\}$. Suppose that α_j is the proper subset of α_{i+1} other than α_i that has the highest index. Choose a vector v with $v_k = 1$ such that $\varphi_{\alpha_i}(v)\varphi_{\alpha_j}(v) = -1$. Now apply Lemma 2.8 with this v. We have

$$w_{i+1} > (2^{k-1} - 1)w_i + w_j - \sum_{\alpha \subset \alpha_{i+1}, \alpha = \alpha_j, \alpha_i} \varphi_{\alpha_i}(v)\varphi_\alpha(v)w_\alpha.$$

Thus the lemma follows from establishing

$$\sum_{\alpha \subset \alpha_{i+1}, \alpha = \alpha_j, \alpha_i} |w_\alpha| \leq w_j.$$

We divide the sum into those α of size at least 3 and those of size at most 2. To bound the first sum we note that since $w_{l+1} > 3w_l$ when $|\alpha_{l+1}| \geq 3$ and $l < j$, even the sum over all sets of size at least 3 and index less than j is bounded by $w_j/2$.

To bound the second sum we observe that there are at most $k(k-1)/2 + k + 1$ terms and each is bounded the maximal weight of a set of size 2. Now there are at least $k - 2$ sets of size 3 before α_j in the enumeration (this bound is tight for $k = 4$ but very weak otherwise), which implies by induction that w_j is at least 3^{k-1} times the maximal weight of any set of size 2. The inequality $3^{k-1} > k(k-1) + 2k + 2$ valid for $k \geq 4$ concludes this case.

Finally suppose $|\alpha_{i+1}| = |\alpha_i| = k$ and $k \geq 4$. We assume that $\alpha_{i+1} = \{1, 2, \ldots, k-1, k+1\}$ and $\alpha_i = \{1, \ldots, k\}$. Suppose that α_j is the set of highest index that appears in the sum of Lemma 2.7. Choose a vector v with $v_k = v_{k+1} = 1$ such that $\varphi_{\alpha_i}(v)\varphi_{\alpha_j}(v) = -1$. Now apply Lemma 2.7 with this v. The analysis is similar to the last case. This finishes the proof of the lemma. □

All that remains to prove Theorem 2.4 is a simple calculation. Let α_{i_0} be the last set of size 2 in our ordering. Then

$$w_{n-1} \geq \prod_{|\alpha_i|>2} (2^{|\alpha_i|-1} - 1)w_{i_0} > 2^{\sum_{i=1}^{n-1} |\alpha_i|-1} \prod_{|\alpha_i|\geq 2} (1 - 2^{1-|\alpha_i|}),$$

since $w_{i_0} > 1$ and the two factors introduced when $|\alpha_i| = 2$ cancel each other.

The first factor equals $2^{nm/2+1-n}$. This follows because the average size of a subset of $[m]$ is $m/2$, and there are n such sets. The extra "$1 - n$" is the result of -1 inside the summation. To estimate the second factor, we use the fact that when $0 < x < \frac{1}{2}$, then $(1 - x) > e^{-2x}$. Hence

$$\prod_{|\alpha_i|\geq 2} (1 - 2^{1-|\alpha_i|}) > e^{-\sum_{k=2}^{n} \binom{m}{k} 2^{2-k}} \geq e^{-4(1+\frac{1}{2})^m} = e^{-4n^\beta},$$

and the theorem follows. □

Note that by slightly extra work, the constant in front of n^β can be reduced to any value greater than 2. In fact, if we were willing to obtain an extra term, we could in fact obtain the value 2. This would be achieved by using an inequality of the type $1 - x \geq e^{-x-cx^2}$ for an appropriate constant c. We do not think this is of great concern unless the upper bound is improved.

If we use the full strength of Lemma 2.9, then we can actually strengthen Theorem 2.5 to apply to all weights.

THEOREM 2.10. *Assume that n is a power of 2 and*

$$F(f) = \operatorname{sign}\left(\sum_{j=0}^{n-1} w_j f(j)\right),$$

where w_j are integers. Then for $n \geq 8$ and all j, we have $|w_j| \geq (1/2n)e^{-4n^\beta}2^{(n\log n)/2-n}$, where $\beta = \log(\frac{3}{2})$.

Proof. Use Lemma 2.3 to obtain a corresponding expansion

$$\sum_{i=0}^{n-1} w_i'(f, \varphi_{\alpha_i}),$$

where

$$w_i' = \frac{1}{n}\sum_{j=0}^{n-1} w_j \varphi_{\alpha_i}(j).$$

We know by Lemma 2.9 that $w_{n-1}' \geq (2^{m-1}-1)w_{n-2}' \geq (2^{m-1}-1)(2^{m-2}-1)w_{n-3}' \cdots$, where $n = 2^m$. By Lemma 2.3,

$$w_j = \sum_{i=0}^{n-1} w_i' \varphi_{\alpha_i}(j),$$

and for $n \geq 8$,

$$\left|\sum_{i=0}^{n-2} w_i' \varphi_{\alpha_i}(j)\right| \leq \frac{1}{2}w_{n-1}'.$$

Since

$$w_{n-1}' \geq \frac{1}{n}e^{-4n^\beta}2^{(n\log n)/2-n},$$

the theorem follows. □

3. Recalling the upper bound. For completeness, let us recall the proof for the upper bound. Let F be any threshold function and let H_0 be a linear function with the following properties:

1. $\operatorname{sign}(H_0(x)) = F(x)$ for $x \in \{-1,1\}^n$.
2. $|H_0(x)| \geq 1$ for $x \in \{-1,1\}^n$.
3. Among H satisfying the above conditions, it maximizes the number of $x \in \{-1,1\}^n$ such that $|H_0(x)| = 1$. If there are several H giving the same number of such points, choose one arbitrarily.

Since F can be represented by a linear threshold, there is some linear function satisfying the first two conditions, and thus there will exist such a linear function H_0.

Let $x^{(i)}$, $i = 1, 2, \ldots, r$ be the set of points in $\{-1, 1\}^n$ with $|H_0(x^{(i)})| = 1$. We claim that H_0 is uniquely determined by the equations

$$H_0(x^{(i)}) = F(x^{(i)}), \qquad i = 1, 2, \ldots, r.$$

Suppose this was not the case. The set of of solutions to these equations would contain linear functions $H_0 + tH_1$ for all rational t and nonzero H_1. There is some point $x^{(0)} \in \{-1, 1\}^n$ such that $H_1(x^{(0)}) \neq 0$. Suppose for concreteness that $H_1(x^{(0)}) > 0$ and that $H_0(x^{(0)}) < -1$. (Note that we cannot have $|H_0(x^{(0)})| = 1$ since we must have $H_1(x^{(i)}) = 0$ for $1 \leq i \leq r$.) Now let t_0 be the minimal $t > 0$ such that $|H_0(x) + tH_1(x)| = 1$ for some $x \notin \{x^{(1)}, x^{(2)}, \ldots, x^{(r)}\}$. There must be such a t_0, since $t = (-1 - H_0(x^{(0)}))/H_1(x^{(0)})$ gives $|H_0(x^{(0)}) + tH_1(x^{(0)})| = 1$. Since $|H_0(x) + tH_1(x)| \geq 1$ for all $t \in [0, t_0]$ and all $x \in \{-1, 1\}^n$, it follows that the two first conditions are satisfied by $H_0 + t_0H_1$. This implies that we violate the maximality condition used to define H_0 and we have reached a contradiction.

By the claim, the coefficients of H_0 can be obtained by solving a linear system of equations where the coefficients and the right-hand side belong to $\{-1, 1\}$. By Cramer's rule this means that each coefficient of H_0 is given by the ratio of two $(n + 1) \times (n + 1)$ (remember that H_0 has a constant coefficient) determinants with entries in $\{-1, 1\}$. Every coefficient has the same denominator and hence we can clear it. By Hadamard's inequality, each absolute value of a determinant of the above type is bounded by $(n+1)^{(n+1)/2}$. Thus F can be realized with integer weights of absolute value at most $(n+1)^{(n+1)/2}$. To obtain a better bound we need the following lemma.

LEMMA 3.1. *The determinant of an $m \times m$ matrix that has entries from the set $\{-1, 1\}$ is divisible by 2^{m-1}.*

Proof. Add the first row to each other row. Now these rows will consist of elements from $\{-2, 0, 2\}$. The determinant of this matrix is clearly divisible by 2^{m-1}. □

Using the lemma and clearing the common factor 2^n gives the following theorem.

THEOREM 3.2. *A threshold function of n variables can be realized with integer weights of size at most $2^{-n}(n + 1)^{(n+1)/2}$.*

4. Final discussion.

When n is a power of two, we have established upper and lower bounds that are only a subexponential factor apart. It is interesting to note that we do not know how to establish such sharp bounds when n is not a power of 2. It is not clear that this is an important problem in determining the true bounds for every n. After all, taking the function F for the largest power of 2 less than n will give fairly good lower bounds. However, it is one of these problems where we do obtain much better bounds for special values of the parameter.

One natural way to try to prove the lower bounds given by Theorem 2.4 is to try to establish that a random threshold function requires large weights. In view of the fact that there are only 2^{n^2} threshold functions, it is not clear that this could succeed. One natural question that arises is how to define a random threshold function.

One definition is to pick a random point $(w_1, w_2, \ldots, w_n)$ uniformly from the real n-dimensional sphere $(\sum_{i=1}^n w_i^2 = 1)$ and then define the random function to be $\operatorname{sign}(\sum_{i=1}^n w_i x_i)$.

It is not difficult to see that with very high probability we can replace w_i by integers with $O(n)$ bits and keep the same function. It is not clear that this is due to a deficiency in the definition or that this is the typical behavior of threshold functions.

Let us finally note that our results extend to the case where the inputs are from the set $\{0, 1, \ldots, a\}$ for $a \geq 2$ (or the more symmetric range $\{-a, 2-a, 4-a, \ldots, a\}$). We can use the same definition of the function and the proof extends essentially word by word. The only part that does not seem to have a counterpart is Lemma 3.1. We thus obtain a lower bound that is roughly a factor a^n stronger, and an upper bound that is a factor $2^n a^{n+1}$ worse.

Acknowledgments. I thank Pekka Orponen for a stimulating conversation on this topic. I am also grateful to Ilan Newman, Ian Parberry, and Sasha Razborov for helpful comments.

REFERENCES

[1] M. GOLDMANN, J. HÅSTAD, AND A. A. RAZBOROV, *Majority gates vs. general weighted threshold gates*, in Proc. 7th Annual IEEE Conference on Structure in Complexity Theory, Boston, MA, 1992, pp. 2–13.

[2] J. HERTZ, R. KROGH, AND A. PALMER, *An Introduction to the Theory of Neural Computation*, Addison-Wesley, Reading, MA, 1991.

[3] P. M. LEWIS AND C. L. COATES, *Threshold Logic*, John Wiley, New York, 1967.

[4] S. MUROGA, *Threshold Logic and its Applications*, Wiley-Interscience, New York, 1971.

[5] P. ORPONEN, *Neural Networks and Complexity Theory*, in Proc. 17th Internat. Sympos. on Mathematical Foundations of Computer Science, Prague, Czechoslovakia, 1992, pp. 50–61.

[6] I. PARBERRY, *The computational and learning complexity of neural networks*, MIT Press, in preparation.

[7] V. ROYCHOWDHURY, K.-Y. SIU, A. ORLISKY, AND T. KAILATH, *A geometric approach to threshold circuit complexity*, in Proc. 4th Ann. Workshop on Computational Learning Theory, Santa Cruz, CA, 1991, pp. 97–111.

[8] D. R. SMITH, *Bounds on the number of threshold functions*, IEEE Trans. Elec. Comput., EC-15 (1966), pp. 368–369.

[9] S. YAJIMA AND T. IBARAKI, *A lower bound on the number of threshold functions*, IEEE Trans. Elec. Comput., EC-14 (1965), pp. 926–929.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 493–498, August 1994

013

DOMINATION, FRACTIONAL DOMINATION, 2-PACKING, AND GRAPH PRODUCTS*

DAVID C. FISHER†

Abstract. Let $P_2(G)$, $\gamma_f(G)$, and $\gamma(G)$ be the *2-packing number*, *fractional domination number*, and *domination number*, respectively, of a graph G. Domke, Hedetniemi, and Laskar [*Congress. Numer.*, 66 (1989), pp. 227–238] showed that $P_2(G) \leq \gamma_f(G) \leq \gamma(G)$. Examples are given with $P_2(G) < \gamma_f(G) = \gamma(G)$ and $P_2(G) = \gamma_f(G) < \gamma(G)$. Let $G \oplus H$ and $G \cdot H$ be the *Cartesian product* and *strong direct product*, respectively, of graphs G and H. For all G and H, it is shown that $P_2(G)P_2(H) \leq P_2(G \cdot H) \leq P_2(G)\gamma_f(H)$ and $\gamma(G)\gamma_f(H) \leq \gamma(G \cdot H) \leq \gamma(G)\gamma(H)$. These relations are also independent. Relations involving $P_2(G \oplus H)$, $\gamma_f(G \oplus H)$, and $\gamma(G \oplus H)$ are examined. An unresolved issue involves a conjecture of Vizing: For all G and H, is $\gamma(G \oplus H) \geq \gamma(G)\gamma(H)$?

Key words. domination, fractional domination, 2-packing, strong direct products, Cartesian products

AMS subject classification. 05C70

Introduction. Let G be a graph. A subset of the nodes of G, S, *dominates* G, if each node in $G - S$ is adjacent to a member of S. The *domination number* of G, $\gamma(G)$, is the minimal cardinality of a dominating set. Domke, Hedetniemi, and Laskar [1] give a fractional version of $\gamma(G)$. A set of nonnegative weights on the nodes of G is a *fractional domination* of G if, for each node a, the weights on a and its neighbors sum to at least one. The *fractional domination number* of G, $\gamma_f(G)$, is the minimal sum of the weights of a fractional domination. A subset S is a *2-packing* of G if, for all $x, y \in S$, the distance between x and y is at least 3. Let the *2-packing number* of G, $P_2(G)$, be the maximal cardinality of a 2-packing. We then have (from [1])

$$P_2(G) \leq \gamma_f(G) \leq \gamma(G). \tag{1}$$

Let G and H be graphs with nodes V_G and V_H and edges E_G and E_H, respectively. Then $G \oplus H$ (the *Cartesian product* of G and H) has nodes (g, h) for all $g \in V_G$ and $h \in V_H$ with an edge between (g_1, h_1) and (g_2, h_2) if and only if $[g_1, g_2] \in E_G$ and $h_1 = h_2$, or $g_1 = g_2$ and $[h_1, h_2] \in E_H$. Also, $G \cdot H$ (the *strong direct product* of G and H) has nodes (g, h) for all $g \in V_G$ and $h \in V_H$ with an edge between (g_1, h_1) and (g_2, h_2) if and only if $g_1 = g_2$ or $[g_1, g_2] \in E_G$, and $h_1 = h_2$ or $[h_1, h_2] \in E_H$ (see Fig. 1). A long-standing conjecture of Vizing [4] is that for all graphs G and H, $\gamma(G \oplus H) \geq \gamma(G)\gamma(H)$. Fisher, Ryan, Domke, and Majumdar [2] give a fractional version of this: $\gamma_f(G \oplus H) \geq \gamma_f(G)\gamma_f(H)$, which follows from a result for strong direct products,

$$\gamma_f(G \cdot H) = \gamma_f(G)\gamma_f(H). \tag{2}$$

Section 1 shows that the inequalities of (1) are independent by giving examples with $P_2(G) < \gamma_f(G) = \gamma(G)$ and $P_2(G) = \gamma_f(G) < \gamma(G)$. Section 2 proves two relations similar to (2): $\gamma(G)\gamma_f(H) \leq \gamma(G \cdot H) \leq \gamma(G)\gamma(H)$ and $P_2(G)P_2(H) \leq P_2(G \cdot H) \leq P_2(G)\gamma_f(H)$. Section 3 summarizes these results in Fig. 7. Relations for $P_2(G \oplus H)$, $\gamma_f(G \oplus H)$, and $\gamma(G \oplus H)$ are also examined.

* Received by the editors July 31, 1991; accepted for publication (in revised form) June 15, 1993.

† Department of Mathematics, University of Colorado at Denver, Campus Box 170, P.O. Box 173364, Denver, Colorado 80217-3364.

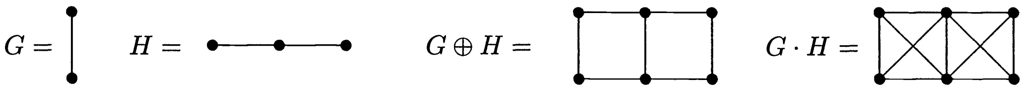


FIG. 1. *The Cartesian product and the strong direct product.*

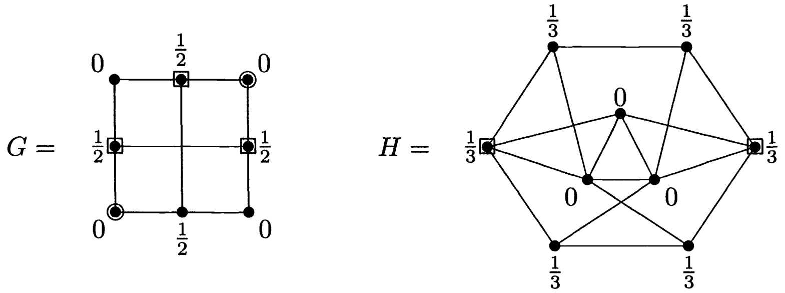


FIG. 2. *The circled nodes are a* 2*-packing of* G*. So* $P_2(G) \geq 2$*. The weights are both a fractional domination and a fractional* 2*-packing of* G*. So* $\gamma_f(G) = 2$*. While the boxed nodes dominate* G*, no two-node set dominates* G*. Hence* $2 = P_2(G) = \gamma_f(G) < \gamma(G) = 3$*. In* H*, the distance between two nodes is at most two. So* $P_2(H) = 1$*. The weights are a fractional* 2*-packing of* H*. So* $\gamma_f(H) \geq 2$*. The boxed nodes dominate* H*. So* $\gamma(H) \leq 2$*. Hence by* (1), $1 = P_2(H) < \gamma_f(H) = \gamma(H) = 2$*.*

1. Extremes for the fractional domination number. Four relations for $P_2(G)$, $\gamma_f(G)$, and $\gamma(G)$ are allowed by (1). Examples with $P_2(G) = \gamma_f(G) = \gamma(G)$ (e.g., K_1) and $P_2(G) < \gamma_f(G) < \gamma(G)$ (e.g., C_4) are easy to find. However, examples with $P_2(G) = \gamma_f(G) < \gamma(G)$ or $P_2(G) < \gamma_f(G) = \gamma(G)$ are surprisingly difficult to find. Fig. 2 gives what seems to be the smallest examples.

2. Bounds on the strong direct product. This section gives lower and upper bounds on $\gamma(G \cdot H)$ and $P_2(G \cdot H)$. For each, examples (Figs. 3–6) are given with the lower bound achieved, but the upper is not, and vice versa. These examples are selected so that at least two integers are within the bounds.

To prove the results, it is easiest to view the graph parameters as optimal values of linear or integer programming problems. For vectors $\mathbf{x}$ and $\mathbf{y}$, let $\mathbf{x} \geq \mathbf{y}$ ($\mathbf{x} \leq \mathbf{y}$) mean $x_i \geq y_i$ ($x_i \leq y_i$) for all i. Let $\mathbf{1}_k$ and $\mathbf{0}_k$ be the k-vectors whose components are all one or zero, respectively. For an m node graph G, let $N(G)$ (the *neighborhood matrix* of G) be the $m \times m$ matrix with $n_{ij} = 1$, if $i = j$ or if nodes i and j are adjacent, and $n_{ij} = 0$, otherwise.

We can redefine $\gamma(G)$ as the value of an integer programming problem

$$(3) \qquad \begin{array}{l} \gamma(G) = \min_{\mathbf{x}} \mathbf{1}_m^T \mathbf{x} \\ \text{subject to } \mathbf{x} \geq \mathbf{0}_m, \quad N(G)\mathbf{x} \geq \mathbf{1}_m \quad \text{and} \quad \mathbf{x} \text{ is integer.} \end{array}$$

A feasible solution to (3) *dominates* G. Similarly, $\gamma_f(G)$ can be redefined as the value of a linear programming problem

$$(4) \qquad \gamma_f(G) = \min_{\mathbf{x}} \mathbf{1}_m^T \mathbf{x} \quad \text{subject to } \mathbf{x} \geq \mathbf{0}_m \quad \text{and} \quad N(G)\mathbf{x} \geq \mathbf{1}_m.$$

A feasible solution to (4) is a *fractional domination* of G. Duality gives another formulation for $\gamma_f(G)$,

$$(5) \qquad \gamma_f(G) = \max_{\mathbf{y}} \mathbf{1}_m^T \mathbf{y} \quad \text{subject to } \mathbf{y} \geq \mathbf{0}_m \quad \text{and} \quad N(G)\mathbf{y} \leq \mathbf{1}_m.$$

A feasible solution to (5) is a *fractional 2-packing* of G. Also, $P_2(G)$ can be redefined as the value of an integer programming problem,

$$(6)\quad P_2(G) = \max_{\mathbf{y}} \mathbf{1}_m^T\mathbf{y} \quad \text{subject to } \mathbf{y} \geq \mathbf{0}_m, \quad N(G)\mathbf{y} \leq \mathbf{1}_m, \quad \text{and} \quad \mathbf{y} \text{ is integer.}$$

A feasible solution to (6) is a *2-packing* of G.

Re-indexing (3) and (6) gives a useful alternative formulation for $\gamma(G \cdot H)$ and $P_2(G \cdot H)$. Let $\mathcal{Z}(m,n)$ be the set of nonnegative integer $m \times n$ matrices. Then

$$(7)\ \gamma(G \cdot H) = \min_{Z} \mathbf{1}_m^T Z \mathbf{1}_n \quad \text{subject to } N(G)ZN(H) \geq \mathbf{1}_m\mathbf{1}_n^T \quad \text{and} \quad Z \in \mathcal{Z}(m,n),$$

$$(8)\quad P_2(G \cdot H) = \max_{Z} \mathbf{1}_m^T Z \mathbf{1}_n \quad \text{subject to } N(G)ZN(H) \leq \mathbf{1}_m\mathbf{1}_n^T \quad \text{and} \quad Z \in \mathcal{Z}(m,n).$$

THEOREM 1. *For all graphs G and H, $\gamma(G)\gamma_f(H) \leq \gamma(G \cdot H) \leq \gamma(G)\gamma(H)$.*

Proof. For the lower bound, let Z be an optimal solution to (7). Then $N(G)ZN(H) \geq \mathbf{1}_m\mathbf{1}_n^T$. Isolating the jth column gives

$$(N(G)ZN(H))_j \geq \mathbf{1}_m.$$

Because $(N(G)ZN(H))_j = N(G)(ZN(H))_j$, we have that $(ZN(H))_j$ dominates G. Thus $\mathbf{1}_m^T(ZN(H))_j \geq \gamma(G)$, and hence $\mathbf{1}_m^T ZN(H) \geq \gamma(G)\mathbf{1}_n^T$. Transposing both sides and rearranging we obtain

$$N(H)\left(\gamma(G)^{-1}Z^T\mathbf{1}_m\right) \geq \mathbf{1}_n.$$

Thus, $\gamma(G)^{-1}Z^T\mathbf{1}_m$ is a fractional domination of H giving

$$\gamma_f(H) \leq \mathbf{1}_n^T\gamma(G)^{-1}Z^T\mathbf{1}_m = \gamma(G)^{-1}\mathbf{1}_m^T Z\mathbf{1}_n = \gamma(G)^{-1}\gamma(G \cdot H).$$

For the upper bound, let $\mathbf{x}$ and $\mathbf{y}$ be dominations of G and H, respectively. Then

$$N(G)\mathbf{x}\mathbf{y}^T N(H) = (N(G)\mathbf{x})(N(H)\mathbf{y})^T \geq \mathbf{1}_m\mathbf{1}_n^T.$$

So $\mathbf{x}\mathbf{y}^T$ is a feasible solution of (7) giving

$$\gamma(G \cdot H) \leq \mathbf{1}_m^T\mathbf{x}\mathbf{y}^T\mathbf{1}_n = (\mathbf{1}_m^T\mathbf{x})(\mathbf{1}_n^T\mathbf{y}) = \gamma(G)\gamma(H). \qquad \square$$

The exactness of the bounds in Theorem 1 are independent. Figs. 3 and 4 give examples with $\gamma(G)\gamma_f(H) = \gamma(G \cdot H) < \gamma(G)\gamma(H)$ and $\gamma(G)\gamma_f(H) < \gamma(G \cdot H) = \gamma(G)\gamma(H)$.

THEOREM 2. *For all graphs G and H, $P_2(G)P_2(H) \leq P_2(G \cdot H) \leq P_2(G)\gamma_f(H)$.*

Proof. For the lower bound, let $\mathbf{x}$ and $\mathbf{y}$ be 2-packings of G and H, respectively. Then

$$N(G)\mathbf{x}\mathbf{y}^T N(H) = (N(G)\mathbf{x})(N(H)\mathbf{y})^T \leq \mathbf{1}_m\mathbf{1}_n^T.$$

So $\mathbf{x}\mathbf{y}^T$ is a feasible solution of (8) giving

$$P_2(G \cdot H) \geq \mathbf{1}_m^T\mathbf{x}\mathbf{y}^T\mathbf{1}_n = (\mathbf{1}_m^T\mathbf{x})(\mathbf{1}_n^T\mathbf{y}) = P_2(G)P_2(H).$$

For the upper bound, let Z be an optimal solution to (8). Then $N(G)ZN(H) \leq \mathbf{1}_m\mathbf{1}_n^T$. Isolating the jth column, we obtain

$$(N(G)ZN(H))_j \leq \mathbf{1}_m.$$

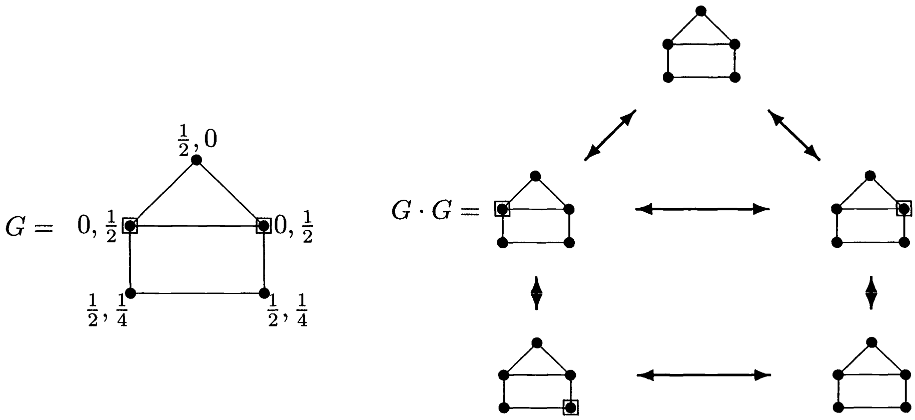


FIG. 3. *The left weights are a fractional* 2*-packing on* G. *The right weights are a fractional domination. Since both sum to* $3/2$, $\gamma_f(G) = 3/2$. *The* 2 *boxed nodes dominate* G. *So* $\gamma(G) = 2$. *The* 3 *boxed nodes dominate* $G \cdot G$. *So, Theorem* 1 *gives* $3 = \gamma(G)\gamma_f(G) = \gamma(G \cdot G) < \gamma(G)\gamma(G) = 4$.

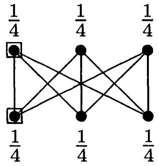


FIG. 4. *The weights on* G *are both a fractional* 2*-packing and a fractional domination of* G. *So* $\gamma_f(G) = 3/2$. *The* 2 *boxed nodes dominate* G. *Thus* $\gamma(G) = 2$. *The graph* $G \cdot G$ *can be divided into two halves, each with* 3 *isolated copies of* G. *Suppose a* 3*-node set* S *dominates* $G \cdot G$. *Since a node can dominate only* 4 *nodes in its half, each half must have at least one member of* S. *In the half with two members, there is a copy of* G *that has no members of* S. *Since the members in its own half cannot dominate its nodes and at most* 4 *nodes can be dominated by the members in the other half,* S *cannot dominate all nodes in this copy. So,* 3 *nodes cannot dominate* $G \cdot G$. *Theorem* 1 *gives* $3 = \gamma(G)\gamma_f(G) < \gamma(G \cdot G) = \gamma(G)\gamma(G) = 4$.

Because $(N(G)ZN(H))_j = N(G)(ZN(H))_j$, we have that $(ZN(H))_j$ is a 2-packing of G. Thus $\mathbf{1}_m^T(ZN(H))_j \le P_2(G)$, and hence $\mathbf{1}_m^T ZN(H) \le P_2(G)\mathbf{1}_n^T$. Transposing both sides and rearranging we obtain

$$N(H)\left(P_2(G)^{-1}Z^T\mathbf{1}_m\right) \le \mathbf{1}_n.$$

Thus, $P_2(G)^{-1}Z^T\mathbf{1}_m$ is a fractional domination of H giving that

$$\gamma_f(H) \ge \mathbf{1}_n^T P_2(G)^{-1}Z^T\mathbf{1}_m = P_2(G)^{-1}\mathbf{1}_m^T Z\mathbf{1}_n \ge P_2(G)^{-1}P_2(G \cdot H). \qquad \square$$

The exactness of the bounds in Theorem 2 are independent. Figs. 5 and 6 give examples with $P_2(G)P_2(H) = P_2(G \cdot H) < P_2(G)\gamma_f(H)$ and $P_2(G)P_2(H) < P_2(G \cdot H) = P_2(G)\gamma_f(H)$.

3. Summary. Equations (1) and (2) and the theorems form a system of inequalities. *Where do* $P_2(G \oplus H)$, $\gamma_f(G \oplus H)$, *and* $\gamma(G \oplus H)$ *fit in*? Since $N(G \oplus H) \le N(G \cdot H)$, replacing $G \cdot H$ by $G \oplus H$ strengthens the constraints in (3) and (4), while weakening those in (5) and (6). So $P_2(G \oplus H) \ge P_2(G \cdot H)$, $\gamma_f(G \oplus H) \ge \gamma_f(G \cdot H)$, and $\gamma(G \oplus H) \ge \gamma(G \cdot H)$. Fig. 7 summarizes the inequalities.

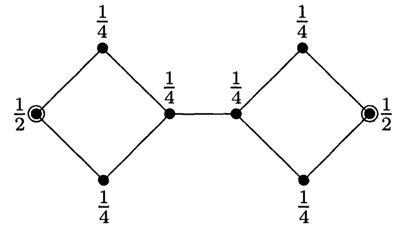


FIG. 5. *The weights are both a fractional domination and a fractional 2-packing of G. Thus, $\gamma_f(G) = 5/2$. The 2 circled nodes are a 2-packing of G. Thus, $P_2(G) = 2$. The graph $G \cdot G$ can be divided into two halves isomorphic to $C_4 \cdot G$. Since Theorem 2 gives $P_2(C_4 \cdot G) \le P_2(C_4)\gamma_f(G) = 5/2$, each half has at most 2 nodes in a 2-packing of $G \cdot G$. So Theorem 2 gives $4 = P_2(G)P_2(G) = P_2(G \cdot G) < P_2(G)\gamma_f(G) = 5$.*

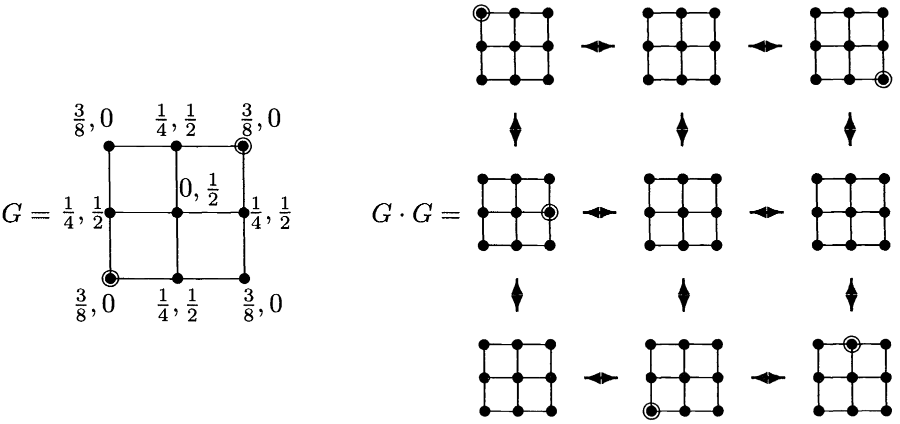

$\frac{3}{8}, 0$
$\frac{1}{4}, \frac{1}{2}$
$\frac{3}{8}, 0$
$G = \frac{1}{4}, \frac{1}{2}$
$0, \frac{1}{2}$
$\frac{1}{4}, \frac{1}{2}$
$\frac{3}{8}, 0$
$\frac{1}{4}, \frac{1}{2}$
$\frac{3}{8}, 0$
$G \cdot G =$

FIG. 6. *The left weights are a fractional 2-packing of G. The right weights are a fractional domination. Since both sum to $5/2$, $\gamma_f(G) = 5/2$. The 2 circled nodes are a 2-packing of G. So $P_2(G) = 2$. The 5 circled nodes are a 2-packing of $G \cdot G$. So Theorem 2 gives $4 = P_2(G)P_2(G) < P_2(G \cdot G) = P_2(G)\gamma_f(G) = 5$.*

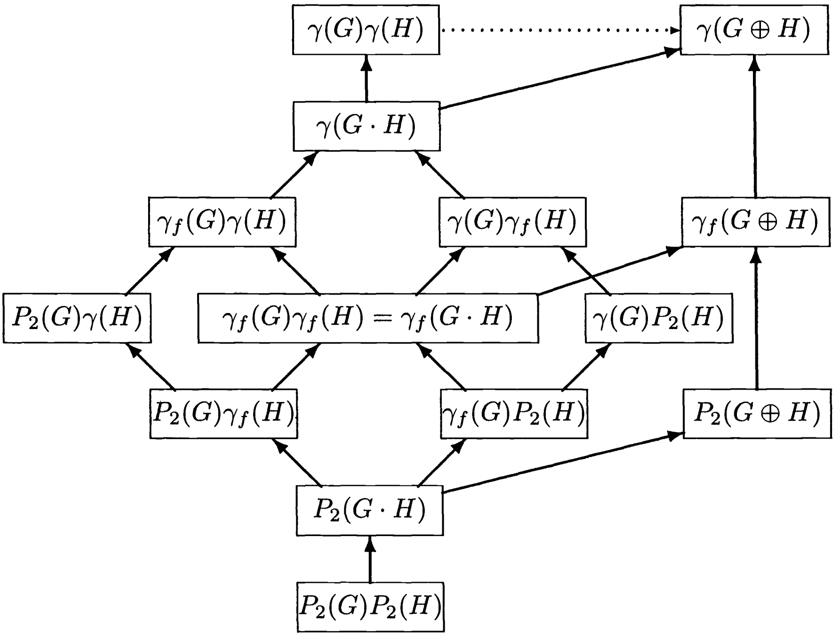


FIG. 7. The big picture. *The arcs indicate that for all graphs G and H, the quantity at the head of the arc is greater than or equal to the one at the tail. The dotted arc is Vizing's conjecture. Quantities without a path between them are independent in that examples exists where each is larger than the other. Many corollaries are possible, such as a result of Jacobson and Kinch* [3] *that for all G and H, $\gamma(G \oplus H) \geq \gamma(G)P_2(H)$.*

Does Fig. 7 show all possible inequalities? Since $P_2(C_4) < \gamma_f(C_4) < \gamma(C_4)$, we have $P_2(K_1 \oplus C_4) < P_2(K_1)\gamma_f(C_4)$ and $\gamma_f(K_1 \oplus C_4) < P_2(K_1)\gamma(C_4)$. Further, let P_3 be the 3-node graph with 2 edges. Then $2 = P_2(P_3 \oplus P_3) > \gamma(P_3)\gamma(P_3) = 1$. So no relations other than those in Fig. 7 are possible (except for Vizing's conjecture).

REFERENCES

[1] G. DOMKE, S. T. HEDETNIEMI, AND R. LASKAR, *Fractional domination in graphs*, Congress. Numer., 66 (1989), pp. 227–238.

[2] D. C. FISHER, J. RYAN, G. S. DOMKE, AND A. MAJUMDAR, *Fractional domination of strong direct products*, Discrete Appl. Math., to appear.

[3] M. S. JACOBSON AND L. F. KINCH, *On the domination number of a graph* II: *Trees*, J. Graph Theory, 10 (1986), pp. 97–106.

[4] V. G. VIZING, *The Cartesian product of graphs*, Vychisl. Sistemy, 9 (1963), pp. 30–43.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 499–512, August 1994

014

NORMAL AND SELF-DUAL NORMAL BASES FROM FACTORIZATION OF $cx^{q+1} + dx^q - ax - b$*

IAN F. BLAKE[†], SHUHONG GAO[‡§], AND RONALD C. MULLIN[‡]

Abstract. The present paper is interested in a family of normal bases, considered by Sidel'nikov [*Math. USSR-Sb.*, 61 (1988), pp. 485–494], with the property that all the elements in a basis can be obtained from one element by repeatedly applying to it a linear fractional function of the form $\varphi(x) = (ax + b)/(cx + d)$, $a, b, c, d \in F_q$. Sidel'nikov proved that the products for such a basis $\{\alpha_i\}$ are of the form $\alpha_i\alpha_j = e_{i-j}\alpha_i + e_{j-i}\alpha_j + \gamma$, $i = j$, where $e_k, \gamma \in F_q$. It is shown that every such basis can be formed by the roots of an irreducible factor of $F(x) = cx^{q+1} + dx^q - ax - b$. The following are constructed: (a) a normal basis of F_{q^n} over F_q with complexity at most $3n - 2$ for each divisor n of $q - 1$ and for $n = p$, where p is the characteristic of F_q; (b) a self-dual normal basis of F_{q^n} over F_q for $n = p$ and for each odd divisor n of $q - 1$ or $q + 1$. When $n = p$, the self-dual normal basis constructed of F_{q^p} over F_q also has complexity at most $3p - 2$. In all cases, the irreducible polynomials and the multiplication tables are given explicitly.

Key words. finite field, irreducible polynomial, normal basis

AMS subject classifications. 11T30, 11T06

1. Introduction.

Let $N = \{\alpha_0, \alpha_1, \ldots, \alpha_{n-1}\}$ be a normal basis of F_{q^n} over F_q, with $\alpha_i = \alpha^{q^i}, 0 \le i \le n - 1$, where q is a prime power p^m, with p a prime and $m \ge 1$. The multiplication of elements in F_{q^n} is uniquely determined by the n products $\alpha_0\alpha_i = \sum_{j=0}^{n-1} t_{ij}\alpha_j, t_{ij} \in F_q$. The $n \times n$ matrix $T = (t_{ij})$ is called the multiplication table of N. As in [6], the number of nonzero elements in T is called the complexity of the normal basis N, denoted by C_N. In hardware and software implementations of finite field arithmetic, normal bases of low complexity offer considerable advantages. In [6] it is proved that $C_N \ge 2n - 1$. When the lower bound is reached, N is called an optimal normal basis of F_{q^n} over F_q. Two families of optimal normal bases are constructed in [6], and in [3] it is proved that these two families are essentially all the optimal normal bases in finite fields. Some normal bases of low complexity are constructed in [1]. A normal basis with the smallest complexity, if no optimal normal bases exist, is called a minimal normal basis.

This paper is interested in a family of normal bases, considered by Sidel'nikov [8], with the property that all the elements in a basis can be obtained from one element by repeatedly applying to it a linear fractional function of the form $\varphi(x) = (ax+b)/(cx+d)$, $a, b, c, d \in F_q$. Sidel'nikov proved that the products for such a basis $\{\alpha_i\}$ are of the form $\alpha_i\alpha_j = e_{i-j}\alpha_i + e_{j-i}\alpha_j + \gamma$, $i \neq j$, where $e_k, \gamma \in F_q$. We show that every such basis can be formed by the roots of an irreducible factor of $F(x) = cx^{q+1}+dx^q-ax-b$. We construct a normal basis of F_{q^n} over F_q, with complexity at most $3n - 2$ for each divisor n of $q - 1$ and for $n = p$, where p is the characteristic of F_q, and a self-dual normal basis of F_{q^n} over F_q for $n = p$ and for each odd divisor n of $q-1$ or $q+1$. When $n = p$, the self-dual normal basis of F_{q^p} constructed over F_q also has complexity of at most $3p - 2$. In all cases, we give the irreducible polynomials and the multiplication

* Received by the editors September 24, 1992; accepted for publication (in revised form) June 22, 1993.

[†] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1 (ifblake@claude2.uwaterloo.ca).

[‡] Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1 (rcmullin@watmath.uwaterloo.ca).

[§] Present address, Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4 (sgao@cs.toronto.edu).

tables explicitly. For this purpose, some properties of linear fractional functions and the complete factorization of $F(x)$ are discussed in §§2 and 3, respectively.

2. On linear fractional functions. In this section, we discuss some properties of the linear fractional function $\varphi(x) = (ax+b)/(cx+d)$, with $a, b, c, d \in F_q$ and $ad - bc \neq 0$. It is easy to see that $\varphi(x)$ defines a permutation on $F_q \cup \{\infty\}$, where

$$\frac{a\infty + b}{c\infty + d} := \frac{a}{c}, \quad \text{if } c \neq 0,$$
$$\frac{a\infty + b}{c\infty + d} := \infty, \quad \text{if } ad \neq 0, c = 0,$$
$$\frac{a}{0} := \infty, \quad \text{if } a \neq 0.$$

Actually, $\varphi(x)$ induces a permutation on $F_{q^n} \cup \{\infty\}$ for any $n \geq 1$. The inverse of $\varphi(x)$ is $\varphi^{-1}(x) = (-dx+b)/(cx-a)$.

For any two linear fractional functions φ and ψ, the composition $\varphi\psi$, defined as $\varphi\psi(x) = \varphi(\psi(x))$, is still a linear fractional function. It is well known that all the linear fractional functions over F_q form a group under composition, and that it is isomorphic to the projective general linear group $PGL(2, q)$. The order of φ is the smallest positive integer t such that $\varphi^t(x) = x$, i.e., φ^t is the identity map.

For our purpose, we deal with a linear fractional function $\varphi(x) = (ax+b)/(cx+d)$ with $c \neq 0$. The fixed points of $\varphi(x)$ satisfy

$$cx^2 - (a-d)x - b = 0. \tag{1}$$

The following two lemmas are easily checked.

LEMMA 2.1. *Let $\varphi(x) = ax + b$, with $a \neq 0, 1$, be a linear mapping. Then*

$$\varphi = h^{-1}\psi h,$$

where $\psi(x) = ax$ and $h(x) = x + b/(a-1)$.

LEMMA 2.2. *Let $\varphi(x) = (ax+b)/(cx+d)$, with $c \neq 0$ and $ad - bc \neq 0$. Let $\Delta = (a-d)^2 + 4bc$. Then*

$$\varphi = h^{-1}\psi h,$$

where $h(x)$ and $\psi(x)$ are defined as follows:

(a) *When $\Delta = 0$, let x_0 be the only solution of* (1) *in F_q, that is, x_0 satisfies $cx_0^2 = -b$ and $2cx_0 = a - d$. Then $h(x) = (a/c - x_0)/(x - x_0)$ and $\psi(x) = x + 1$.*

(b) *When $\Delta \neq 0$, let x_0, x_1 be the two solutions of* (1) *in F_{q^2}, and let $\xi = (a - cx_0)/(a - cx_1)$. Then*

$$h(x) = \frac{x - x_0}{x - x_1}, \qquad \psi(x) = \xi x.$$

The order of φ is now easy to determine: it is equal to the order of ψ. If ψ is of the form $x + 1$, then the order of ψ is equal to the additive order p of 1 in F_q, where p is the characteristic of F_q. If ψ is of the form ξx, then the order of ψ is equal to the multiplicative order of ξ. In case (b) of Lemma 2.2, if Δ is a quadratic residue in F_q, then $x_0, x_1 \in F_q$, and $\xi \in F_q$. Hence $\xi^{q-1} = 1$, and the order of ξ is a divisor of $q - 1$. If Δ is a quadratic nonresidue in F_q, then $x_0, x_1 \in F_{q^2} \setminus F_q$ and $x_0^q = x_1, x_1^q = x_0$. Thus $\xi^q = ((a - cx_0)/(a - cx_1))^q = (a - cx_0^q)/(a - cx_1^q) = (a - cx_1)/(a - cx_0) = 1/\xi$.

So $\xi^{q+1} = 1$, and the order of ξ divides $q+1$. Therefore, the order of φ is always a divisor of p, $q-1$, or $q+1$.

LEMMA 2.3. *Let $a, b, c, d \in F_q$ with $c \neq 0$, and $ad - bc \neq 0$. Let $\varphi(x) = (ax+b)/(cx+d)$, with order t. Then, for $1 \leq i \leq t-1$,*

$$\varphi^i(x) = \frac{e_i x + b/c}{x - e_{t-i}}, \qquad e_i + e_{t-i} = \frac{a-d}{c}, \tag{2}$$

where $e_1 = a/c$ and $e_{i+1} = \varphi(e_i)$ for $i = 1, \ldots, t-2$.

Proof. It is routine to prove by induction on i that there exist $e_i, f_i \in F_q$ with $e_1 = a/c$, $f_1 = d/c$ such that

$$\varphi^i(x) = \frac{e_i x + b/c}{x + f_i},$$

and

$$e_i - f_i = \frac{a-d}{c}, \qquad e_i = \varphi(e_{i-1}),$$

for $i = 1, \ldots, t-1$, where $e_0 = \infty$. Note that

$$\frac{e_{t-i}x + b/c}{x + f_{t-i}} = \varphi^{t-i}(x) = \varphi^{-i}(x) = (\varphi^i)^{-1}(x) = \frac{-f_i x + b/c}{x - e_i}.$$

We see that $f_i = -e_{t-i}$. This completes the proof. □

LEMMA 2.4. *With the same notation as in Lemma 2.3, we have that*

$$\sum_{j=1}^{t-1} e_j = \begin{cases} (t-1)(a-d)/(2c), & \text{if } p \neq 2, \\ a/c = d/c, & \text{if } p = 2 \text{ and } t = 2, \\ (a-d)/c, & \text{if } p = 2 \text{ and } t \equiv 3 \bmod 4, \\ 0, & \text{if } p = 2 \text{ and } t \equiv 1 \bmod 4, \end{cases} \tag{3}$$

where p is the characteristic of F_q.

Proof. We consider two cases according to the type of $\varphi(x)$.

Case I. $\Delta = (a-d)^2 + 4bc = 0$. Then $t = p$ and, by Lemma 2.2, $\varphi(x) = h^{-1}\psi h(x)$, where

$$\psi(x) = x + 1, \quad h(x) = \frac{a/c - x_0}{x - x_0}, \quad h^{-1}(x) = x_0 + \frac{a/c - x_0}{x},$$

with x_0 satisfying $2cx_0 = a - d$ and $cx_0^2 = -b$. Note that $\psi^i(x) = x + i$. We have that

$$\begin{aligned} \varphi^i(x) &= h^{-1}\psi^i h(x) \\ &= h^{-1}\left(\frac{a/c - x_0}{x - x_0} + i\right) \\ &= \frac{(a/c - x_0 - ix_0)x - ix_0^2}{ix + (a/c - x_0 - ix_0)}. \end{aligned}$$

So

$$e_i = \frac{a/c - x_0}{i} + x_0, \quad \text{for } 1 \leq i \leq t-1.$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{p-1} e_i &= (p-1)x_0 + (a/c - x_0)\sum_{i=1}^{p-1} i^{-1}\\ &= (p-1)x_0 + (a/c - x_0)\sum_{i=1}^{p-1} i\\ &= \begin{cases} (p-1)x_0 = (t-1)(a-d)/(2c), & \text{if } p \neq 2,\\ a/c = d/c, & \text{if } p = 2.\end{cases}\end{aligned}$$

Case II. $\Delta = (a-d)^2 + 4bc \neq 0$. In this case, the order t of $\varphi(x)$ is a factor of $q-1$ or $q+1$. So $t \not\equiv 0 \bmod p$. By Lemma 2.2, $\varphi(x) = h^{-1}\psi h(x)$, where

$$h(x) = \frac{x - x_0}{x - x_1}, \quad \psi(x) = \xi x, \quad \xi = \frac{a/c - x_0}{a/c - x_1},$$

with $x_0 + x_1 = (a-d)/c$ and $x_0 x_1 = -b/c$. Note that $h^{-1}(x) = (x_1 x - x_0)/(x-1)$ and $\psi^i(x) = \xi^i x$. We have that

$$\begin{aligned}\varphi^i(x) &= h^{-1}\psi^i h(x)\\ &= h^{-1}\left(\xi^i \frac{x - x_0}{x - x_1}\right)\\ &= \frac{(x_1\xi^i - x_0)x - x_0x_1(\xi^i - 1)}{(\xi^i - 1)x + x_1 - x_0\xi^i}.\end{aligned}$$

So

$$e_i = \frac{x_1\xi^i - x_0}{\xi^i - 1} = x_1 + \frac{x_1 - x_0}{\xi^i - 1}, \quad \text{for } 1 \le i \le t-1,$$

and

$$\sum_{i=1}^{t-1} e_i = (t-1)x_1 + (x_0 - x_1)\sum_{i=1}^{t-1}\frac{1}{1-\xi^i}.$$

Because ξ is a tth primitive root of unity, we have that

$$\prod_{i=1}^{t-1}(x - \xi^i) = (x^t - 1)/(x-1) = x^{t-1} + x^{t-2} + \cdots + x + 1. \tag{4}$$

Letting $x = 1$ in (4), we obtain the following:

$$\prod_{i=1}^{t-1}(1 - \xi^i) = t. \tag{5}$$

Taking derivatives with respect to x on both sides of (4), we have that

$$\prod_{i=1}^{t-1}(x - \xi^i)\left(\sum_{i=1}^{t-1}\frac{1}{x - \xi^i}\right) = (t-1)x^{t-2} + (t-2)x^{t-3} + \cdots + 2x + 1. \tag{6}$$

Letting $x = 1$ in (6), we see that

$$\sum_{i=1}^{t-1} \frac{1}{1-\xi^i} = \left(\sum_{i=1}^{t-1} i\right)/t = \begin{cases} (t-1)/2, & \text{if } p \neq 2, \\ 1, & \text{if } p = 2 \text{ and } t \equiv 3 \bmod 4, \\ 0, & \text{if } p = 2 \text{ and } t \equiv 1 \bmod 4, \end{cases}$$

(note that t is odd when $p = 2$). Therefore,

$$\sum_{i=1}^{t-1} e_i = \begin{cases} ((t-1)/2)(x_0 + x_1) = (t-1)(a-d)/(2c), & \text{if } p \neq 2, \\ x_0 - x_1 = (a-d)/c, & \text{if } p = 2 \text{ and } t \equiv 3 \bmod 4, \\ 0, & \text{if } p = 2 \text{ and } t \equiv 1 \bmod 4. \end{cases}$$

This completes the proof. □

The following theorem is proved by Sidel'nikov [8, Thm. 2].

THEOREM 2.5. *Let $a, b, c, d \in F_q$, with $c \neq 0$ and $ad - bc \neq 0$. Let θ be a root of $F(x) = cx^{q+1} + dx^q - ax - b$ in some extension field of F_q not fixed by $\varphi(x) = (ax + b)/(cx + d)$, whose order is assumed to be t. Then*

$$\theta, \varphi(\theta), \ldots, \varphi^{t-1}(\theta)$$

are linearly independent over F_q, if $\sum_{i=0}^{t-1} \varphi^i(\theta) \neq 0$.

This theorem indicates that if we can factor $F(x)$, then we can obtain normal bases over F_q. The factorization of $F(x)$ is discussed in the next section.

3. Factorization of $cx^{q+1} + dx^q - ax - b$. The complete factorization of $F(x) = cx^{q+1} + dx^q - ax - b$, $a, b, c, d \in F_q$, into irreducible factors was established by Ore [7, pp. 264–270], using his theory of linearized polynomials. In this section, we briefly discuss how this can be done without resorting to linearized polynomials. For the detail, the reader is referred to [2]. To exclude the trivial cases, we assume that $ad - bc \neq 0$. Let $\varphi(x) = (ax + b)/(cx + d)$ be the linear fractional function associated with $F(x)$. As noted in §2, $\varphi(x)$ induces a permutation on $F_{q^n} \cup \{\infty\}$, for any $n \geq 1$. We assume that the order of φ is t in this section.

Let θ be a root of $F(x) = (cx + d)x^q - (ax + b)$. Then

$$\theta^q = \frac{a\theta + b}{c\theta + d} = \varphi(\theta).$$

Note that

$$\theta^{q^2} = (\varphi(\theta))^q = \varphi(\theta^q) = \varphi(\varphi(\theta)) = \varphi^2(\theta).$$

By induction, we see that $\theta^{q^i} = \varphi^i(\theta)$, $i \geq 0$. So

$$\theta, \varphi(\theta), \ldots, \varphi^{t-1}(\theta) \tag{7}$$

are all the conjugates of θ over F_q. If θ is a fixed point of $\varphi(x)$, then $\theta \in F_q$, and $x - \theta$ is a factor of $F(x)$. If θ is not a fixed point of $\varphi(x)$, then, by Theorem 2.5, the elements of (7) are distinct and θ is of degree t over F_q. In the latter case, the minimal polynomial of θ over F_q is an irreducible factor of $F(x)$ of degree t. So an irreducible factor of $F(x)$ is either linear or of degree t. We first deal with two special cases.

THEOREM 3.1. *Let $\xi \in F_q \setminus \{0\}$, with multiplicative order t. Then the following factorization over F_q is complete:*

$$x^{q-1} - \xi = \prod_{j=1}^{(q-1)/t} (x^t - \beta_j),$$

where β_j are all the $(q-1)/t$ distinct roots of $x^{(q-1)/t}-\xi$ in F_q.

Proof. Let θ be a root of $x^{q-1}-\xi$ in some extension field of F_q. Then $\theta^{q^i}=\theta\xi^i, i\geq 1$. All the distinct conjugates of θ over F_q are $\theta,\theta\xi,\ldots,\theta\xi^{t-1}$. The minimal polynomial of θ over F_q is

$$\prod_{i=0}^{t-1}(x-\theta\xi^i)=x^t-\theta^t,$$

which divides $x^{q-1}-\xi$. This means that any irreducible factor of $x^{q-1}-\xi$ is of the form $x^t-\beta$, where $\beta\in F_q$. Note that $x^t-\beta$ divides $x^{q-1}-\xi$, if and only if β is a root of $x^{(q-1)/t}-\xi$. This completes the proof. □

THEOREM 3.2. *For $x^q-(x+b)$ with $b\in F_q^*$, the following factorization over F_q is complete:*

$$x^q-(x+b)=\prod_{j=1}^{q/p}(x^p-b^{p-1}x-b^p\beta_j), \tag{8}$$

where β_j are the distinct elements of F_q with $Tr_{q/p}(\beta_j)=1$ and p is the characteristic of F_q.

Proof. Let θ be a root of $F(x)=x^q-(x+b)$. Then $\theta^{q^i}=\theta+ib, i\geq 1$. So the conjugates of θ over F_q are $\theta,\theta+b,\ldots,\theta+(p-1)b$. The minimal polynomial of θ over F_q is

$$\begin{aligned}\prod_{i=0}^{p-1}[x-(\theta+ib)]&=b^p\prod_{i=0}^{p-1}\left[\frac{x-\theta}{b}-i\right]\\&=b^p\left[\left(\frac{x-\theta}{b}\right)^p-\frac{x-\theta}{b}\right]\\&=x^p-b^{p-1}x+\theta(b^{p-1}-\theta^{p-1}).\end{aligned}$$

Hence, an irreducible factor of $x^q-(x+b)$ is of the form

$$x^p-b^{p-1}x-\beta,\quad \beta\in F_q. \tag{9}$$

Let γ be a root of (9) in some extension field of F_q. Then we have that

$$\left(\frac{\gamma}{b}\right)^{p^i}-\left(\frac{\gamma}{b}\right)^{p^{i-1}}=\left(\frac{\beta}{b^p}\right)^{p^{i-1}},\quad 1\leq i\leq m, \tag{10}$$

where $q=p^m$. Summing (10) yields

$$\gamma^{p^m}-\gamma=b\mathrm{Tr}_{q/p}\left(\frac{\beta}{b^p}\right).$$

Consequently, (9) divides $F(x)=x^{p^m}-x-b$ if and only if $\mathrm{Tr}_{q/p}(\beta/b^p)=1$. Note that there are $q/p=p^{m-1}$ elements β in F_q with trace 1, and the proof is completed. □

In general, we show that the factorization of $F(x)$ can be reduced to factoring x^q-x-1, $x^{q-1}-\xi$, or $x^{q+1}-\xi$. Let $\varphi=h^{-1}\psi h$, as in Lemmas 2.1 and 2.2. For any root θ of $F(x)$ that is not fixed by φ, we have that

$$h(\theta^q)=\psi(h(\theta)). \tag{11}$$

If Δ is a quadratic residue in F_q, then $h(\theta^q) = (h(\theta))^q$. Thus $\eta = h(\theta)$ is a root of $x^q - x - 1$ or $x^q - \xi x = x(x^{q-1} - \xi)$, when $\psi(x) = x + 1$ or $\psi(x) = \xi x, \xi \in F_q$. So, by the factorization of $x^q - x - 1$ and $x^{q-1} - \xi$ as in Theorems 3.1 and 3.2, we obtain the factorization of $F(x)$ as follows.

THEOREM 3.3. *For $a, b \in F_q$ with $a \neq 0, 1$, the following factorization over F_q is complete*:

$$x^q - (ax + b) = \left(x - \frac{b}{a-1}\right) \prod_{j=1}^{(q-1)/t} \left(\left(x - \frac{b}{a-1}\right)^t - \beta_j\right),$$

where t is the multiplicative order of a, and β_j are all the $(q-1)/t$ distinct roots of $x^{(q-1)/t} - a$.

THEOREM 3.4. *For $a, b, c, d \in F_q$, with $c \neq 0$, $ad-bc \neq 0$, and $\Delta = (a-d)^2+4bc = 0$, the following factorization over F_q is complete*:

$$\begin{aligned}&(cx + d)x^q - (ax + b) \\ &\quad = (x - x_0) \prod_{j=1}^{q/p} \left[(x - x_0)^p + \frac{1}{\beta_j}(a/c - x_0)(x - x_0)^{p-1} - \frac{1}{\beta_j}(a/c - x_0)^p\right],\end{aligned}$$

where $x_0 \in F_q$ is the unique solution of (1), *and β_j are all the q/p distinct elements of F_q with $Tr_{q/p}(\beta_j) = 1$.*

THEOREM 3.5. *For $a, b, c, d \in F_q$, with $c \neq 0$, $ad-bc \neq 0$, and $\Delta = (a-d)^2+4bc \neq 0$ being a quadratic residue in F_q, the following factorization over F_q is complete*:

$$\begin{aligned}&(cx + d)x^q - (ax + b) \\ &\quad = (x - x_0)(x - x_1) \prod_{j=1}^{(q-1)/t} \frac{1}{1 - \beta_j}[(x - x_0)^t - \beta_j(x - x_1)^t],\end{aligned}$$

where $x_0, x_1 \in F_q$ are the two distinct roots of (1), *t is the multiplicative order of $\xi = (a - cx_0)/(a - cx_1)$, and β_j are all the $(q-1)/t$ distinct roots of $x^{(q-1)/t} - \xi$ in F_q.*

If Δ is not a quadratic residue in F_q, the situation is a little more complicated, as in this case $x_0, x_1, \xi \notin F_q$. Noting that $x_0^q = x_1$ and $x_1^q = x_0$, we have that $h(\theta^q) = (1/h(\theta))^q$. Equation (11) implies that $\eta = 1/h(\theta)$ is a root of $x^{q+1} - \xi$. So, by factoring $x^{q+1} - \xi$ over F_{q^2}, we can obtain the factorization of $F(x)$ over F_{q^2}. Then by "combining" these factors, we obtain the factorization of $F(x)$ over F_q as in Theorem 3.6.

THEOREM 3.6. *For $a, b, c, d \in F_q$, with $c \neq 0$, $ad-bc \neq 0$, and $\Delta = (a-d)^2+4bc \neq 0$ being a quadratic nonresidue in F_q, the following factorization over F_q is complete*·

$$\begin{aligned}F(x) &= (cx + d)x^q - (ax + b) \\ &= \prod_{j=1}^{(q+1)/t} \frac{1}{1 - \beta_j}[(x - x_0)^t - \beta_j(x - x_1)^t], \qquad (12)\end{aligned}$$

where $x_0, x_1 \in F_{q^2}$ are the two distinct roots of (1), *t is the multiplicative order of $\xi = (a - cx_1)/(a - cx_0)$, and β_j are all the $(q+1)/t$ distinct roots of $x^{(q+1)/t} - \xi$ in F_{q^2}.*

Let $f(x)$ be any nonlinear irreducible factor of $F(x)$ of degree t, and let α be a root of $f(x)$. From the discussion at the beginning of this section, we see that

$\varphi^i(\alpha), i = 0, 1, \ldots, t-1$ are all the roots of $f(x)$ and, by Theorem 2.5, they are linearly independent over F_q if $\mathrm{Tr}(\alpha) \neq 0$. However, $\mathrm{Tr}(\alpha)$ is just the negative of the coefficient of x^{t-1} in $f(x)$. By examining the factors in the above explicit factorizations, we have the following theorem.

THEOREM 3.7. *Let $F(x) = (cx+d)x^q - (ax+b)$, with $a, b, c, d \in F_q$, $c \neq 0$, and $ad - bc \neq 0$. Then a monic nonlinear irreducible factor $f(x)$ of $F(x)$ of degree t has linearly dependent roots over F_q if and only if the coefficient of x^{t-1} in $f(x)$ is zero. The latter happens only if $\Delta = (a-d)^2 + 4bc \neq 0$ and $f(x)$ is of the form*

$$\frac{1}{x_1 - x_0}[x_1(x-x_0)^t - x_0(x-x_1)^t],$$

where x_0 and x_1 are solutions of (1).

This shows that every nonlinear irreducible factor of $F(x)$ with one possible exception, has linearly independent roots.

4. Normal bases. As Theorem 3.7 shows, when $c \neq 0$, the roots of an irreducible nonlinear factor of $F(x)$ form a normal basis over F_q (with the possible exception of one factor). This section is devoted to discussing the properties of these bases. We show how to construct a normal basis of F_{q^n} over F_q with complexity at most $3n-2$ for $n = p$, and for each divisor n of $q-1$. For this purpose, we first compute the multiplication tables of the normal bases formed by the roots of an irreducible factor of $F(x)$.

Without loss of generality, we assume that $F(x) = x^{q+1} + dx^q - ax - b$ with $a, b, d \in F_q$ and $b \neq ad$. Assume that $\varphi(x) = (ax+b)/(x+d)$ has order n and that, by Lemma 2.3, $\varphi^i(x) = (e_i x + b)/(x - e_{n-i})$, with $e_i = \varphi^{i-1}(a)$, $1 \leq i \leq n-1$. Let $f(x)$ be any irreducible nonlinear factor of $F(x)$ and α a root of $f(x)$. Then $f(x)$ has degree n and its roots are the following:

$$\alpha_i = \alpha^{q^i} = \varphi^i(\alpha), \quad i = 0, 1, \ldots, n-1.$$

They form a normal basis of F_{q^n} over F_q if the coefficient of x^{n-1} in $f(x)$ is not zero (or $\mathrm{Tr}(\alpha) \neq 0$), by Theorem 3.7.

THEOREM 4.1. *Let $F(x) = x^{q+1} + dx^q - (ax+b)$, with $a, b, d \in F_q$ and $b \neq ad$. Let $f(x)$ be an irreducible factor of $F(x)$ of degree $n > 1$ and let α be a root of it. Then all the roots of $f(x)$ are the following:*

$$\alpha_i = \alpha^{q^i} = \varphi^i(\alpha), \quad i = 0, 1, \ldots, n-1, \tag{13}$$

where $\varphi(x) = (ax+b)/(x+d)$. If $\tau = \sum_{i=0}^{n-1} \alpha_i$, the negative of the coefficient of x^{n-1} in $f(x)$ is not zero. Then, (13) *form a normal basis of F_{q^n} over F_q such that*

$$\alpha_0 \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-1} \end{pmatrix} = \begin{pmatrix} \tau^* & -e_{n-1} & -e_{n-2} & \cdots & -e_1 \\ e_1 & e_{n-1} & & & \\ e_2 & & e_{n-2} & & \\ \vdots & & & \ddots & \\ e_{n-1} & & & & e_1 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-1} \end{pmatrix} + \begin{pmatrix} b^* \\ b \\ b \\ \vdots \\ b \end{pmatrix} \tag{14}$$

where $e_1 = a$, $e_{i+1} = \varphi(e_i)$ $(i \geq 1)$, $b^* = -b(n-1)$, and $\tau^* = \tau - \epsilon$ with

$$\epsilon = \sum_{i=1}^{n-1} e_i = \begin{cases} (n-1)(a-d)/2, & \text{if } p \neq 2, \\ a = d, & \text{if } p = n = 2, \\ a - d, & \text{if } p = 2 \text{ and } n \equiv 3 \bmod 4, \\ 0, & \text{if } p = 2 \text{ and } n \equiv 1 \bmod 4. \end{cases}$$

Proof. We must only prove (14). By Lemma 2.3, for $i \geq 1$,

$$\alpha_i = \varphi^i(\alpha) = \frac{e_i\alpha_0 + b}{\alpha_0 - e_{n-i}}.$$

So

$$\alpha_0\alpha_i = e_i\alpha_0 + e_{n-i}\alpha_i + b.$$

For $i = 0$, we have that

$$\alpha_0\alpha_0 = \alpha_0\left(\tau - \sum_{j=1}^{n-1}\alpha_j\right) = \left(\tau - \sum_{j=1}^{n-1} e_j\right)\alpha_0 - \sum_{j=1}^{n-1} e_{n-j}\alpha_j - b(n-1).$$

The theorem follows from Lemma 2.4. □

The next theorem can be viewed as the "converse" of Theorem 4.1.

THEOREM 4.2. *Let $n > 2$ and $\alpha_i = \alpha^{q^i}$, for $0 \leq i \leq n-1$. Suppose that $\{\alpha_i\}$ is a normal basis of F_{q^n} over F_q and satisfies the following*:

(15) $$\alpha_i\alpha_j = a_{ij}\alpha_i + b_{ij}\alpha_j + \gamma_{ij}, \quad \text{for all } 0 \leq i \neq j \leq n-1,$$

where $a_{ij}, b_{ij}, \gamma_{ij} \in F_q$. Then there are constants $\gamma, e_1, e_2, \ldots, e_{n-1} \in F_q$ such that

(a) *$e_i = \varphi(e_{i-1})$, for $2 \leq i \leq n-1$, and*

$$a_{ij} = e_{j-i}, \quad b_{ij} = e_{i-j}, \quad \gamma_{ij} = \gamma, \quad \text{for all } i \neq j,$$

where $\varphi(x) = (e_1x + \gamma)/(x - e_{n-1})$, and the subscripts of e are calculated modulo n;

(b) *The minimal polynomial of α is a factor of $F(x) = x^{q+1} - e_{n-1}x^q - (e_1x + \gamma)$, and thus n must be a factor of p, $q-1$, or $q+1$.*

Proof. Let $e_k = a_{0k}$ and $\gamma_k = \gamma_{0k}$ for $k = 1, 2, \ldots, n-1$. Then

(16) $$\alpha_0\alpha_k = e_k\alpha_0 + b_{0k}\alpha_k + \gamma_k.$$

Raising (16) to the q^{n-k}th power on both sides, we have that

(17) $$\alpha_0\alpha_{n-k} = b_{0k}\alpha_0 + e_k\alpha_{n-k} + \gamma_k.$$

Subtracting (17) from (16), with the k in (16) replaced by $n-k$, gives

(18) $$(e_{n-k} - b_{0k})\alpha_0 + (b_{0\,n-k} - e_k)\alpha_{n-k} + \gamma_{n-k} - \gamma_k = 0.$$

Because $n > 2$ and the α_i's are linearly independent over F_q, (18) implies that

$$b_{0k} = e_{n-k}, \quad \gamma_k = \gamma_{n-k}, \quad 1 \leq k \leq n-1.$$

Therefore,

(19) $$\alpha_0\alpha_k = e_k\alpha_0 + e_{n-k}\alpha_k + \gamma_k, \quad 1 \leq k \leq n-1.$$

Now, for any $i \neq j$, raising (19) to the q^ith power and letting $k = j - i$, we have that

(20) $$\alpha_i\alpha_j = e_{j-i}\alpha_i + e_{i-j}\alpha_j + \gamma_{j-i}.$$

Comparing (20) and (15) gives

$$a_{ij} = e_{j-i}, \quad b_{ij} = e_{i-j}, \quad \gamma_{ij} = \gamma_{j-i}, \tag{21}$$

which proves part of (a).

We prove the remaining part of (a) together with (b). To this purpose, note that a special case of (20) is

$$\alpha_i \alpha_{i+1} = e_{n-1}\alpha_{i+1} + e_1\alpha_i + \gamma_1, \quad 0 \le i < n-1,$$

or

$$\alpha_{i+1} = \frac{e_1\alpha_i + \gamma_1}{\alpha_i - e_{n-1}} = \varphi(\alpha_i), \quad 0 \le i < n-1, \tag{22}$$

where $\varphi(x) = (e_1 x + \gamma)/(x - e_{n-1})$ with $\gamma = \gamma_1$. So, by induction on i, we see that $\alpha_i = \varphi^i(\alpha_0) = \varphi^i(\alpha), 0 \le i \le n-1$. We know, by Lemma 2.3, that

$$\varphi^i(x) = (a_i x + \gamma)/(x - a_{n-i}), \quad 0 \le i \le n-1,$$

where $a_i = \varphi(a_{i-1})$, for $i \ge 1$, and $a_1 = e_1$. Thus (22) implies that

$$\alpha_i = \frac{a_i\alpha_0 + \gamma}{\alpha_0 - a_{n-i}},$$

i.e.,

$$\alpha_0\alpha_i = a_i\alpha_0 + a_{n-i}\alpha_i + \gamma. \tag{23}$$

Comparing (23) to (19), we have that

$$e_i = a_i, \quad e_{n-i} = a_{n-i}, \quad \gamma_i = \gamma.$$

This proves (a). For (b), note that $\alpha_1 = \alpha^q$, and that (19) with $k = 1$ means α is a root of $F(x) = x^{q+1} - e_{n-1}x^q - e_1 x - \gamma$. Therefore, the minimal polynomial of α divides $F(x)$. This completes the proof. □

THEOREM 4.3. *For every $a, \beta \in F_q^*$ with $Tr_{q/p}(\beta) = 1$,*

$$x^p - \frac{1}{\beta}ax^{p-1} - \frac{1}{\beta}a^p \tag{24}$$

is irreducible over F_q, and its roots form a normal basis of F_{q^p} over F_q with complexity at most $3p-2$. The multiplication table is

$$\begin{pmatrix} \tau^* & -e_{p-1} & -e_{p-2} & \cdots & -e_1 \\ e_1 & e_{p-1} & & & \\ e_2 & & e_{p-2} & & \\ \vdots & & & \ddots & \\ e_{p-1} & & & & e_1 \end{pmatrix}, \tag{25}$$

where $e_1 = a$, $e_{i+1} = \varphi(e_i)$ for $i \ge 1$, $\varphi(x) = ax/(x+a)$, and $\tau^* = a/\beta$ if $p \ne 2$ or $a/\beta - a$ if $p = 2$.

Proof. Let $F(x) = (x+a)x^q - ax$ and $\varphi(x) = ax/(x+a)$. Then, $F(x)$ satisfies the conditions of Theorem 3.4 with $b = 0, c = 1, d = a, \Delta = 0$, and $x_0 = 0$. So, (24) is an irreducible factor of $F(x)$. As the coefficient of x^{p-1} in (24) is $-a/\beta \ne 0$, by Theorem 4.1, the roots of (24) form a normal basis and its multiplication table is (25). The complexity is obviously at most $3p-2$. □

THEOREM 4.4. *Let n be any factor of $q-1$. Let $\beta \in F_q$ with multiplicative order t such that* $\gcd(n,(q-1)/t)=1$ *and let $a=\beta^{(q-1)/n}$. Then*

$$x^n - \beta(x-a+1)^n \tag{26}$$

is irreducible over F_q and its roots form a normal basis of F_{q^n} over F_q of complexity at most $3n-2$. The multiplication table is

$$\begin{pmatrix} \tau^* & -e_{n-1} & -e_{n-2} & \cdots & -e_1 \\ e_1 & e_{n-1} & & & \\ e_2 & & e_{n-2} & & \\ \vdots & & & \ddots & \\ e_{n-1} & & & & e_1 \end{pmatrix}, \tag{27}$$

where $e_1=a$, $e_{i+1}=\varphi(e_i)$ $(i\geq 1)$, $\varphi(x)=ax/(x+1)$, and $\tau^*=-n(a-1)\beta/(1-\beta)-\epsilon$, with ϵ specified as in Theorem 4.1 (*with $d=1$*).

Proof. It is easy to see that a has multiplicative order n. Then $\varphi(x)=ax/(x+1)$ has $x_0=0$ and $x_1=a-1$ as fixed points, and $\xi=(a-x_0)/(a-x_1)=a$ has order n. So φ has order n. Note that β is a root of $x^{(q-1)/n}-a$. By Theorem 3.5, the polynomial (26) is an irreducible factor of $F(x)=x^{q+1}+x^q-ax$. Note that the coefficient of x^{n-1} in (26) is $n(a-1)\neq 0$. By Theorem 4.1 (with $b=0, d=1$), the roots of (26) form a normal basis of F_{q^n} over F_q, and its multiplication table is (27). The complexity is obviously at most $3n-2$. □

Table 1 is the result of a computer search for the minimal complexity of normal bases. It indicates that when $n|(q-1)$, the minimal complexity is often $3n-3$ or $3n-2$. This indicates that the normal bases constructed in Theorems 4.3 and 4.4 often have complexity very close to the minimal complexity. In the table, † indicates that the minimal complexity is $3n-2$, and ⋆ indicates optimal complexity, i.e., $2n-1$. Other minimal values are of the form $3n-3$.

TABLE 1

q	5	7	7	11	11	13	13	17	19
n	4	3	6	5	10	3	4	4	3
min	9	6	16†	12	28†	6	7⋆	7⋆	6

5. Self-dual normal bases.

A basis $B=\{\beta_0,\beta_1,\ldots,\beta_{n-1}\}$ is called a dual basis of $A=\{\alpha_0,\alpha_1,\ldots,\alpha_{n-1}\}$ if $\mathrm{Tr}(\alpha_i\beta_j)=\delta_{ij}=0$ for $i\neq j$, and 1 for $i=j$, where Tr is the trace function of F_{q^n} into F_q, defined as $\mathrm{Tr}(\alpha)=\alpha+\alpha^q+\cdots+\alpha^{q^{n-1}}\in F_q$, $\alpha\in F_{q^n}$. We can prove that, for each basis A of F_{q^n} over F_q, there is a unique dual basis. Also, if A is normal then so is its dual. If the dual basis of A coincides with A, then A is called a self-dual basis, that is, a basis $A=\{\alpha_i\}$ is called self-dual if $\mathrm{Tr}(\alpha_i\alpha_j)=\delta_{ij}$. Lempel and Weinberger [5] proved the following theorem.

THEOREM 5.1. *A self-dual normal basis of F_{q^n} over F_q exists if and only if one of the following conditions is satisfied*:

(a) *q is even and n is not a multiple of* 4;

(b) *both q and n are odd.*

Later, Jungnickel, Menezes, and Vanstone [4] determined the total number of self-dual bases and self-dual normal bases of F_{q^n} over F_q.

However, the proofs of these results are not constructive. In this section, we construct a self-dual normal basis of F_{q^n} over F_q for every n in the following cases:

(a) $n = p$, the characteristic of F_q;

(b) $n|(q-1)$ and n is odd;

(c) $n|(q+1)$ and n is odd.

THEOREM 5.2. *Let $N = \{\alpha_0, \alpha_1, \cdots, \alpha_{n-1}\}$, with $\alpha_i = \alpha^{q^i}$, be a normal basis of F_{q^n} over F_q satisfying*

$$\alpha_i \alpha_j = e_{j-i}\alpha_i + e_{i-j}\alpha_j + \gamma, \quad \textit{for all } i \neq j,$$

where $e_1, e_2, \ldots, e_{n-1}, \gamma \in F_q$. Let $\tau = Tr_{q^n/q}(\alpha)$ and $\lambda = -(e_1 + e_{n-1}) - n\gamma/\tau$. Then

$$\left\{ \frac{1}{\tau(\tau + n\lambda)}(\alpha_i + \lambda) : \quad i = 0, 1, \ldots, n-1 \right\}$$

is the dual basis of N.

Proof. Note that, for $i \neq j$,

$$\begin{aligned}\mathrm{Tr}_{q^n/q}(\alpha_i(\alpha_j + \lambda)) &= \mathrm{Tr}_{q^n/q}(\lambda\alpha_i + e_{j-i}\alpha_i + e_{i-j}\alpha_j + \gamma)\\ &= \lambda\tau + e_{j-i}\tau + e_{i-j}\tau + n\gamma\\ &= \tau(\lambda + e_1 + e_{n-1}) + n\gamma\\ &= 0,\end{aligned}$$

and

$$\begin{aligned}\mathrm{Tr}_{q^n/q}(\alpha_i(\alpha_i + \lambda)) &= \mathrm{Tr}\left(\alpha_i\left(\tau + \lambda - \sum_{j=i}\alpha_j\right)\right)\\ &= \mathrm{Tr}\left(\alpha_i\left(\tau + n\lambda - \sum_{j=i}(\alpha_j + \lambda)\right)\right)\\ &= \mathrm{Tr}(\alpha_i)(\tau + n\lambda) - \sum_{j=i}\mathrm{Tr}(\alpha_i(\alpha_j + \lambda))\\ &= \tau(\tau + n\lambda).\end{aligned}$$

The result is proved. □

We now proceed to determine when the roots of an irreducible factor of $F(x) = x^{q+1} + dx^q - ax - b$ form a self-dual normal basis. Let $\{\alpha_0, \alpha_1, \ldots, \alpha_{n-1}\}$ be a normal basis generated by a root α of $F(x)$ with $\alpha_i = \alpha^{q^i}$, and let $\tau = \mathrm{Tr}_{q^n|q}(\alpha)$. By Theorem 4.1 and Lemma 2.3, we have, for $i \neq 0$, that

$$\begin{aligned}\mathrm{Tr}_{q^n/q}(\alpha_0\alpha_i) &= e_i\mathrm{Tr}(\alpha_0) + e_{n-i}\mathrm{Tr}(\alpha_i) + nb\\ &= \tau(e_i + e_{n-i}) + nb\\ &= \tau(a-d) + nb, \qquad (28)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Tr}_{q^n/q}(\alpha_0\alpha_0) &= \tau(\tau - \epsilon) - \tau\epsilon - nb(n-1)\\ &= \begin{cases} \tau^2, & \text{if } p = 2,\\ \tau^2 - (n-1)(\tau(a-d) + nb), & \text{if } p \neq 2. \end{cases} \qquad (29)\end{aligned}$$

Therefore, α generates a self-dual normal basis if $\tau = \mathrm{Tr}(\alpha) = 1$ and $(a-d)+nb = 0$. By examining the irreducible factors in Theorems 3.4–3.6, we find that these two conditions can be satisfied. More explicitly, we have the following three results.

THEOREM 5.3. *For any* $\beta \in F_q^*$ *with* $\mathrm{Tr}_{q/p}(\beta) = 1$,

$$x^p - x^{p-1} - \beta^{p-1} \tag{30}$$

is irreducible over F_q, *and its roots form a self-dual normal basis of* F_{q^p} *over* F_q *with complexity at most* $3p-2$. *The multiplication table is* (25) *where* $e_1 = \beta$, $e_{i+1} = \varphi(e_i)$ $(i \geq 1)$, $\varphi(x) = \beta x/(x+\beta)$, *and* $\tau^* = 1$ *if* $p \neq 2$ *or* $\tau^* = 1-\beta$ *if* $p = 2$.

Proof. Let $F(x) = (x+\beta)x^q - \beta x$. Then, by Theorem 3.4, the polynomial (30) is an irreducible factor of $F(x)$ (where $b = 0$, $c = 1$, $d = a = \beta$, $x_0 = 0$ and $\beta_j = \beta$). Since $a - d = b = 0$ and $\tau = 1$ in (28) and (29), the roots of (30) form a self-dual normal basis. Its multiplication table is (25), by Theorem 4.1. □

THEOREM 5.4. *Let* n *be an odd factor of* $q-1$ *and* $\xi \in F_q$ *of multiplicative order* n. *Then there exists* $u \in F_q$, *such that* $(u^2)^{(q-1)/n} = \xi$. *Let* $x_0 = (1+u)/n$ *and* $x_1 = (1+u)/(nu)$. *Then the monic polynomial*

$$\frac{1}{1-u^2}[(x-x_0)^n - u^2(x-x_1)^n] \tag{31}$$

is irreducible over F_q *and its roots form a self-dual normal basis of* F_{q^n} *over* F_q. *The multiplication table is* (14) *with* $a = (x_0 - \xi x_1)/(1-\xi)$, $b = -x_0x_1$, $d = a - (x_0 + x_1)$, *and* $\tau = 1$.

Proof. We first prove that there exists at least one root of $x^{(q-1)/n} - \xi$ that is a quadratic residue in F_q. Let ζ be a primitive element in F_q. Let t be an odd factor of $q-1$ such that $n|t$ and $\gcd(n, (q-1)/t) = 1$. Then $\zeta_0 = \zeta^{(q-1)/t}$ is a tth primitive root of unity. Since t is odd, ζ_0^2 is also a tth primitive root of unity. Let $d = t/n$. Then there is an integer i such that $(\zeta_0^2)^{id} = \xi$, that is,

$$(\zeta^{(q-1)/t})^{2id} = (\zeta^{2i})^{(q-1)/n} = \xi.$$

So ζ^{2i} is a root of $x^{(q-1)/n} - \xi$ and is a quadratic residue in F_q. Therefore, we can take $u = \zeta^i$.

Now, by applying Theorem 3.5, we see that (31) is an irreducible factor of $F(x) = (x+d)x^q - (ax+b)$. The negative of the coefficient of x^{n-1} in (31) is

$$\tau = \frac{n(x_0 - u^2x_1)}{1-u^2} = 1.$$

By Theorem 4.1, the roots of (31) form a normal basis of F_{q^n} over F_q with the claimed multiplication table. Note that

$$a - d = x_0 + x_1 = \frac{(u+1)}{n} + \frac{u+1}{nu} = \frac{(u+1)^2}{nu} = nx_0x_1 = -nb,$$

that is, $\tau(a-d) + nb = 0$. It follows from (28) and (29) that the roots of (31) form a self-dual normal basis. □

THEOREM 5.5. *Let* n *be an odd factor of* $q+1$, *and let* $\xi \in F_{q^2}$ *be a root of* $x^{q+1} - 1$ *with multiplicative order* n. *Then there is a root* u *of* $x^{q+1} - 1$ *such that* $(u^2)^{(q+1)/n} = \xi$. *Let* $x_0 = (1+u)/n$ *and* $x_1 = (1+u)/(nu)$. *Then*

$$\frac{1}{1-u^2}[(x-x_0)^n - u^2(x-x_1)^n] \tag{32}$$

is in $F_q[x]$ and is irreducible over F_q, with its roots forming a self-dual normal basis of F_{q^n} over F_q. The multiplication table is (14) *with $a = (x_1 - \xi x_0)/(1-\xi)$, $b = -x_0 x_1$, $d = a - (x_0 + x_1)$, and $\tau = 1$.*

Proof. The proof of the existence of u is similar to that in the proof of Theorem 5.4, by taking ζ to be a $(q+1)$th primitive root of unity in F_{q^2}. We next prove that $a, b, d \in F_q$ and (32) is in $F_q[x]$. Note that ξ, u, and u^2 are all $(q+1)$th roots of unity, and we have $\xi^q = 1/\xi$, $u^q = 1/u$, and $(u^2)^q = 1/u^2$. Thus $x_0^q = x_1$ and $x_1^q = x_0$. So $a^q = a$, $b^q = b$, and $d^q = d$, that is, $a, b, d \in F_q$. Denote the polynomial (32) by $\phi(x)$ and note that

$$\begin{aligned}(\phi(x))^q &= \frac{1}{1-(u^2)^q}[(x^q - x_0^q)^n - (u^2)^q(x^q - x_1^q)^n] \\ &= \frac{1}{1-1/u^2}[(x^q - x_1)^n - 1/u^2(x^q - x_0)^n] \\ &= \phi(x^q).\end{aligned}$$

We see that the coefficients of $\phi(x)$ are in F_q.

To prove that (32) is irreducible over F_q, we apply Theorem 3.6. It is easy to confirm that, with a, b, d as defined in Theorem 5.5, x_0 and x_1 are the two distinct solutions of (1) with $c = 1$ and $(a - x_1)/(a - x_0) = \xi$, which is of order n. Now, since u^2 is assumed to be a solution of $x^{(q+1)/q} - \xi$, it follows from Theorem 3.6 that (32) is an irreducible factor of $F(x) = (x+d)x^q - (ax+b)$.

As the coefficient of x^{n-1} in (32) is $(-nx_0 + nu^2x_1)/(1-u^2) = -1$, the trace of any root of (32) is $\tau = 1$. It is easy to confirm that $\tau(a-d) + nb = 0$. It follows from (28) and (29) that the roots of (32) form a self-dual normal basis. The multiplication table follows from Theorem 4.1. □

REFERENCES

[1] D. W. Ash, I. F. Blake, and S. A. Vanstone, *Low complexity normal bases*, Discrete Appl. Math., 25 (1989), pp. 191–210.

[2] I. F. Blake, S. Gao, and R. C. Mullin, *Factorization of $cx^{q+1} + dx^q - ax - b$ and normal bases over $GF(q)$*, Research report CORR91-26, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, 1991.

[3] S. Gao and H. W. Lenstra, Jr., *Optimal normal bases*, Designs, Codes and Cryptography, 2 (1992), pp. 315–323.

[4] D. Jungnickel, A. J. Menezes, and S. A. Vanstone, *On the number of self-dual bases of $GF(q^m)$ over $GF(q)$*, Proc. Amer. Math. Soc., 109 (1990), pp. 23–29.

[5] A. Lempel and M. J. Weinberger, *Self-complementary normal bases in finite fields*, SIAM J. Discrete Math., 1 (1988), pp. 193–198.

[6] R. C. Mullin, I. M. Onyszchuk, S. A. Vanstone, and R. M. Wilson, *Optimal normal bases in $GF(p^n)$*, Discrete Appl. Math., 22 (1988/1989), pp. 149-161.

[7] O. Ore, *Contributions to the theory of finite fields*, Trans. Amer. Math. Soc., 36 (1934), pp. 243–274.

[8] V. M. Sidel'nikov, *On normal bases of a finite field*, Math. USSR-Sb., 61 (1988), pp. 485–494.

SIAM J. DISCRETE MATH.
Vol. 7, No. 3, pp. 513–530, August 1994

015

ROUTING PERMUTATIONS ON GRAPHS VIA MATCHINGS*

NOGA ALON[†], F. R. K. CHUNG[‡], AND R. L. GRAHAM[§]

Abstract. A class of routing problems on connected graphs G is considered. Initially, each vertex v of G is occupied by a "pebble" that has a unique destination $\pi(v)$ in G (so that π is a permutation of the vertices of G). It is required that all the pebbles be routed to their respective destinations by performing a sequence of moves of the following type: A disjoint set of edges is selected, and the pebbles at each edge's endpoints are interchanged. The problem of interest is to minimize the number of steps required for any possible permutation π.

This paper investigates this routing problem for a variety of graphs G, including trees, complete graphs, hypercubes, Cartesian products of graphs, expander graphs, and Cayley graphs. In addition, this routing problem is related to certain network flow problems, and to several graph invariants including diameter, eigenvalues, and expansion coefficients.

Key words. eigenvalues, diameters, expanders

AMS subject classifications. 05C, 68

1. Introduction.

Routing problems on graphs arise naturally in a variety of guises, such as the study of communicating processes on networks, data flow on parallel computers, and the analysis of routing algorithms on VLSI chips. A simple (though fundamental) problem of this type is the following. Suppose we are given a connected graph $G = (V, E)$, where V and E represent the vertex and edge sets, respectively, of G. We denote the cardinality $|V|$ of V by n. Initially, each vertex v of G is occupied by a unique marker or "pebble" p. To each pebble p is associated a destination vertex $\pi(v) \in V$, so that distinct pebbles have distinct destinations. Pebbles can be moved to different vertices of G according to the following basic procedure: At each step a disjoint collection of edges of G is selected and the pebbles at each edge's two endpoints are interchanged. Our goal is to move or "route" the pebbles to their respective destinations in a minimum number of steps.

We imagine the steps occurring at discrete times, and we let $p_v(t) \in V$ denote the location of the pebble with initial position v at time $t = 0, 1, 2, \ldots$. Thus, for any t, the set $\{p_v(t) : v \in V\}$ is just a permutation of V. We denote our target permutation that takes v to $\pi(v)$, $v \in V$, by π. Define $rt(G, \pi)$ to be the minimum possible number of steps required to achieve π. Finally, define $rt(G)$, the *routing number* of G, by

$$rt(G) = \max_{\pi} rt(G, \pi),$$

where π ranges over all destination permutations on G. (Sometimes we also call π a routing assignment.)

In more algebraic terms, the problem is simply to determine for G the largest number of terms $\tau = (u_1 v_1)(u_2 v_2) \cdots (u_r v_r)$ ever required to represent any permutation in the symmetric group on $n = |V|$ symbols, where each permutation τ consists of a product of disjoint transpositions $(u_k v_k)$ with all pairs $\{u_k, v_k\}$ required to be edges of G.

* Received by the editors August 31, 1992; accepted for publication (in revised form) June 3, 1993.

[†] Bell Communications Research, Morristown, New Jersey 07960, and Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel.

[‡] Bell Communications Research, Morristown, New Jersey 07960.

[§] AT&T Bell Laboratories, Murray Hill, New Jersey 07974, and Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

To see that $rt(G)$ always exists, let us restrict our attention to some spanning subtree T of G. It is clear that if p has a destination that is a leaf of T, then we can first route p to its destination u, and then complete the routing on $T \setminus \{u\}$ by induction.

In this paper, we investigate routing on a variety of graphs. These include trees, complete graphs, complete bipartite graphs, hypercubes, Cartesian products of graphs, Cayley graphs, and expander graphs. We also consider a related continuous version of the routing problem, the so-called flow problem, which is of independent interest. Furthermore, we relate the routing problem on a given graph to several invariants of it, including its diameter, its resistance, and its expansion coefficients and eigenvalues.

2. General bounds on $rt(G)$. To begin, an obvious lower bound on $rt(G)$ is the following:

$$rt(G) \geq \operatorname{diam}(G), \tag{1}$$

where $\operatorname{diam}(G)$ denotes the diameter of G, i.e., the number of edges in a longest path in G. It would be interesting (but probably difficult) to characterize graphs for which this equality holds.

Suppose C is a cutset of vertices, and let A and B be subsets of V separated by the removal of C. Then

$$rt(G) \geq \frac{2}{|C|} \min(|A|, |B|). \tag{2}$$

This follows by considering the permutation π that maps all pebbles starting in $|A|$ into $|B|$ (where we assume, without loss of generality, that $|A| \leq |B|$). All pebbles in A (i.e., those p_i with $p_i(0) \in A$) must pass through some vertex v of C, and it takes two steps for p_i to pass through v: one to move it from A onto v, and one to move it from v into B (which exchanges it with some pebble from B).

Almost the same argument applies if C is a cutset of edges of G, giving the following similar bound:

$$rt(G) \geq \frac{2}{|C|} \min(|A|, |B|) - 1.$$

This is tight for paths of even length.

Let $\mu(G)$ denote the size of a maximum matching in G. For a routing assignment π, define $D(G, \pi)$ by

$$D(G, \pi) := \sum_{v} d_G(v, \pi(v)),$$

where d_G is the usual (path-) metric on G. Then, setting

$$D(G) := \max_{\pi} D(G, \pi),$$

we have the bound

$$rt(G) \geq \frac{D(G)}{2\mu(G)}.$$

This can be seen by noting that $D(G)$ can only be decreased by at most $2\mu(G)$ at each step.

Since for any spanning subgraph H of G we have

$$rt(G) \leq rt(H),$$

then $rt(G)$ is bounded above by $rt(T)$ for any spanning subtree of G. For any graph G on n vertices, this last quantity is less than $3n$, by Theorem 1, below. We next consider the routing number of trees.

3. Trees. Let $T(n)$ denote some arbitrary fixed tree on n vertices. The following result gives a reasonably good upper bound on $rt(T(n))$.

THEOREM 1.

$$rt(T(n)) < 3n. \tag{3}$$

Proof. We need the following simple and known fact, which can be easily proved (by induction, for example).

Fact. For any tree T on n vertices, there always exists a vertex z of T (see Fig. 1) such that each subtree T_i formed by removing z (and all incident edges) satisfies

$$|T_i| \leq n/2. \tag{4}$$

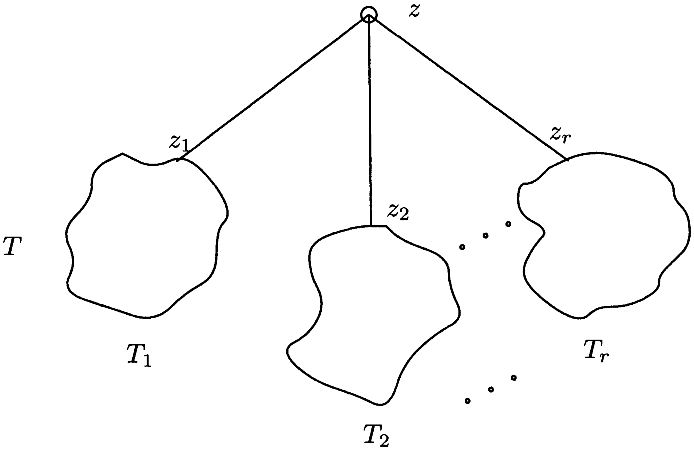


FIG. 1. *Decomposing a tree T.*

The proof of Theorem 1 is by induction on $n = |T|$. Let us apply (4) and let T' denote any one of the subtrees T_i. Consider a pebble $p = p_v(0)$ initially placed on a vertex v of T'. Let us call p *proper* if the destination of p under the routing assignment π belongs to T'; otherwise, call p *improper.* For the special vertex z (the "root"), the pebble $p_z(0)$ is classified as improper.

Our first objective is to move all improper pebbles in (each) T' toward z', the vertex of T' adjacent to z, so that the vertices they occupy form a subtree T'' of T' containing z'.

CLAIM. *The subtree T'' in T' can be formed in at most $|T'|$ steps.*

Proof of Claim. Let $z' = v_1, v_2, \ldots, v_m$ be the vertices on some path M in T'. After the ith step in the process (i.e., after "time i"), there is a certain distribution of pebbles on M. We let $p(v, i)$ denote the pebble occupying vertex v at time i. More generally, we use the index i to denote the value of a parameter at time i. In particular, let $I(i)$ denote the set of *improper* pebbles on M at time i that are further from z' than *some* proper pebble on M (where distance on T is measured by the usual path metric, i.e., the number of edges in the unique path connecting two vertices). Let $x(i)$ denote the set of all improper pebbles in T' that are not in the path M. Also, let $P(i)$ denote the set of *proper* pebbles on M that are closer to z' than some improper pebbles on M. Further, let $C(i)$ denote the set of proper pebbles p on M that are adjacent to an improper pebble on M further from z. Finally, define the function $\phi(i)$, called the *potential*, by

$$(5) \qquad \phi(i) := |I(i)| + |P(i)| + \min\{|P(i)|, |x(i)|\} - |C(i)|.$$

For example, for the distribution (on the path M) shown in Fig. 2 (where • denotes a proper pebble, and ∘ denotes an improper pebble) we have: $|I| = 6$, $|P| = 5$, $|C| = 3$, and (assuming $|x| \geq 5$) $\phi = 13$.

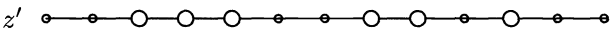


FIG. 2. *Pebbles in a path.*

The algorithm we employ for reaching the desired state is simply a greedy algorithm: Whenever we can interchange an improper and proper pebble so as to bring the improper pebble closer to z', we do it. More specifically, at each step we choose a maximal set of disjoint pairs of this type, and perform the interchanges. We now argue that if we have not yet reached the desired state (i.e., the set of all improper pebbles in T' does not yet span a subtree of T' containing z'), then the potential $\phi(i)$ (computed for some specific path M to be chosen later) must decrease at the next step.

To see this, observe that since our greedy algorithm must eventually terminate, we can find some improper pebble $\hat{p}$ that is moved during the last step. Consider the path $M = (z' = v_1, \ldots, v_m)$, where v_m is the location of $\hat{p}$ at time 0, i.e., $p(v_m, 0) = \hat{p}$. By the definition of our algorithm, no improper pebble is ever moved off of M. On the other hand, it is quite possible that new improper pebbles are moved onto M. Let us denote the pebble distribution on M in terms of alternating blocks of improper and proper pebbles (see Fig. 3). P_j denotes the jth block of proper pebbles (with size $|P_j|$) and I_j denotes the jth block of improper pebbles. Each P_j and I_j, $1 \leq j \leq r$, is nonempty (although I_0 may be empty). By definition, $\phi(i)$ depends only on $x(i)$, P_j and I_j, $1 \leq j \leq r$, and is given by

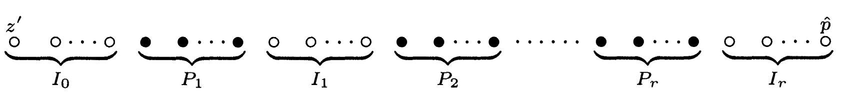


FIG. 3. *Pebbles on M.*

$$\phi(i) = \sum_{j=1}^{r} |I_j| + \sum_{j=1}^{r} |P_j| - r + \min\left(\sum_{j=1}^{r} |P_j|, |x(i)|\right).$$

Now, when we go to time $i + 1$, various changes in M can occur. To begin, the last (i.e., right-most) proper pebble in each P_j is replaced by some improper pebble, either the *first* pebble in I_j or some other improper pebble from outside of M. Observe that if $|I(i)|$ increases during a step, then both $|P(i)|$ and $|x(i)|$ must decrease by at least the same amount. By keeping track of all the possible changes that can occur at the next step, it is not difficult (though it is somewhat tedious) to verify that in all cases, $\phi(i+1) \leq \phi(i) - 1$. We omit the somewhat lengthy details. Since the potential can never exceed, by definition, the number of vertices in T', this completes the proof of the claim. □

The next step in the proof of Theorem 1 is to move each component's improper pebbles to their correct components. With $T_1, \dots, T_r$ denoting the subtrees formed by the removal of the root z, let $\bar{I}(T_j)$ denote the set of improper pebbles in T_j, and let $\bar{P}(T_j)$ denote the set of proper pebbles in T_j. It can easily be shown that using at most three steps, two improper pebbles can be moved to their correct destination components. In fact, if t denotes the largest $|\bar{I}(T_j)|$, then we can move at least $2t$ improper pebbles to their correct destination components in at most $2t+1$ steps.

Following this procedure, we can guarantee that all pebbles are in their correct components (and z is occupied by its proper pebble) in at most

$$\frac{3}{2}\left(\sum_j |\bar{I}(T_j)| - 2t\right) + 2t + 1$$

steps.

Note that by the claim, T'' can be formed (in T_j) in at most $|T_j|$ steps.

Now, since by induction each T_j can now be routed in fewer than $3|T_j|$ steps, then they can all be routed (in parallel) in fewer than $3 \max |T_j|$ steps. Thus, T can be routed in less than

$$\max_j |T_j| + \frac{3}{2}\left(\sum_j |\bar{I}(T_j)|\right) - t + 1 + 3\max |T_j|$$

steps. However, $\max_j |T_j| - t$ does not exceed the number of proper pebbles in the largest subtree. Let y denote this number of proper pebbles. Then, clearly

$$\sum_j |\bar{I}(T_j)| \leq n - 1 - y,$$

and hence the previous quantity is at most

$$y + \frac{3}{2}(n-1-y) + 1 + 3\max |T_j| < \frac{3}{2}n + 3(n/2) = 3n.$$

This completes the induction step; since (3) holds for $n = 2$, Theorem 1 is proved. □

The bound in Theorem 1 can perhaps be improved. For example, it seems clear that one should *not* wait to start moving pebbles across z and routing within T_i's until all improper pebbles in each T_i have been moved close to z_i (i.e., all of these steps can be made in parallel). In fact, the correct value of the constant may be half as large, as suggested by the following conjecture.

Conjecture. For any tree T_n on n vertices,

$$rt(T_n) \leq \left\lfloor \frac{3(n-1)}{2} \right\rfloor. \tag{6}$$

Furthermore, we suspect that the equality can only be achieved when the tree is the star S_n on n vertices.

The fact that equality in (6) holds for S_n was pointed out to us by Goddard [7]. This can be seen as follows.

Denote the center of the star S_n by x. Each cycle $(a_1 \dots a_r)$ of $r \geq 2$ vertices not containing x requires exactly $r+1$ steps to achieve the desired routing. If $(a_1 \dots a_r)$ contains x, then exactly r steps are required. Therefore

$$rt(S_n, \pi) = n - 1 + t,$$

where t denotes the number of cycles of π of size at least 2 that do not contain x. Since $t \leq \lfloor \frac{n-1}{2} \rfloor$, we obtain the desired bound.

For the case where T_n is a path P_n on n vertices, our routing problem reduces to a well-studied problem in parallel sorting networks (see [9] for a comprehensive survey). In this case, it can be shown that $rt(P_n) = n$. In fact, any permutation π on P_n can be sorted in n steps by labelling consecutive edges in P_n as $e_1, e_2, \dots, e_{n-1}$ and only making interchanges with *even* edges e_{2k} on *even* steps and *odd* edges e_{2k+1} on *odd* steps.

4. Complete graphs. Let K_n denote the complete graph on n vertices. In this case, because K_n is so highly connected, the routing number K_n is as small as we could hope.

THEOREM 2. *For the complete graph K_n on $n \geq 3$ vertices,*

$$rt(K_n) = 2. \tag{7}$$

Proof. To see that $rt(K_n) \geq 2$, it is enough to consider the permutation $\pi = (abc)$ consisting of a 3-cycle on K_n. It is clear that such a π cannot be achieved in a single step.

To show that $rt(K_n) \leq 2$, it suffices to show that any *cyclic* permutation can be achieved in two steps, since any permutation π can be factored into disjoint cycles, which can then all be routed in parallel. Thus, let π_m denote the cyclic permutation on $\{1, 2, \dots, m\}$ given by

$$\begin{aligned} &\pi(i) = i - 1, \quad 1 < i \leq m, \\ &\pi(1) = m. \end{aligned}$$

Consider the following two routing steps:

$$S_1 : \ (1, m+1-1)(2, m+1-2) \cdots (i, m+1-i) \cdots,$$

and

$$S_2 : \ (1, m-1)(2, m-2) \cdots (j, m-j) \cdots.$$

We confirm that the composition $S_1 \circ S_2$ sends the following:

$$i \longrightarrow m+1-i \longrightarrow m-(m+1-i) = i-1, \quad i \neq 1,$$

$$1 \longrightarrow m.$$

This map achieves the desired permutation in two steps. Consequently, $rt(K_n) \leq 2$, and the theorem is proved. □

The following result is due to Goddard [7].

THEOREM 3. *For the complete bipartite graph $K_{n,n}$ with $n \geq 3$,*

$$rt(K_{n,n}) = 4. \tag{8}$$

Proof. Suppose $K_{n,n}$ has vertex sets A and B, where the edges are all between A and B. To see that $rt(K_{n,n}) \geq 4$, we consider the permutation $\pi = (a_1 a_2 a_3)$, where a_i's are in A. It is not difficult to show that π cannot be achieved in three steps.

A pebble is said to be an A-pebble if its destination is in A. Otherwise, it is called a B-pebble. In at most one step, we can move all A-pebbles to B and all B-pebbles to A. To prove that $rt(K_{n,n}) \leq 4$, it suffices to show that any cyclic permutation $\pi = (1, 2, \ldots, 2m)$ can be achieved in the following three routing steps:

$$S_1 : (1,2)(3,4)\cdots\left(2\left\lfloor\frac{m}{2}\right\rfloor - 1, 2\left\lfloor\frac{m}{2}\right\rfloor\right)\left(2\left\lfloor\frac{m}{2}\right\rfloor + 2, 2\left\lfloor\frac{m}{2}\right\rfloor + 3\right)\cdots(2m-2, 2m-1),$$
$$S_2 : (1,2m)(3,2m-2)\cdots\left(2\left\lceil\frac{m}{2}\right\rceil - 1, 2\left\lfloor\frac{m}{2}\right\rfloor + 2\right),$$
$$S_3 : (3,2m)\cdots\left(2\left\lfloor\frac{m}{2}\right\rfloor + 1, 2\left\lceil\frac{m}{2}\right\rceil + 2\right).$$

This proves the theorem. □

More generally, it is not difficult to show that for a general complete bipartite graph $K_{m,n}$, $m \leq n$, we have

$$rt(K_{m,n}) \leq 2\left\lceil\frac{n}{m}\right\rceil + 2, \tag{9}$$

since in at most two steps, m pebbles (in fact, B-pebbles, as defined above,) can be routed to their destinations.

5. Cartesian products. For graphs $G = (V, E)$, $G' = (V', E')$, we define the Cartesian product graph $G \times G'$ to be the graph with vertex set $V \times V' = \{(v, v') \mid v \in V, v' \in V'\}$ and with $(u, u')(v, v')$ an edge of $G \times G'$ if and only if either $u = v$, $u'v' \in E'$, or $u' = v'$, $uv \in E$. Thus, the n-cube Q^n is just the Cartesian product of K_2 with itself n times.

The following theorem can be traced back to the early work of Beněs [5]. It was also proved by Baumslag and Annexstein [4].

THEOREM 4. *We have*

$$rt(G \times G') \leq 2rt(G) + rt(G'). \tag{10}$$

Note that since $G \times G'$ and $G' \times G$ are isomorphic graphs, then (10) can be written in the symmetric form

$$rt(G \times G') \leq \min\{2rt(G) + rt(G'),\ 2rt(G') + rt(G)\}. \tag{10'}$$

We briefly describe the proof of Theorem 4 here. We can picture $G \times G'$ as an array $V \times V'$, with each row spanning a copy of G' and each column spanning a copy of G as illustrated in Fig. 4. To route in $G \times G'$, we:

(i) route in columns (copies of G); then
(ii) route in rows (copies of G'); then
(iii) route in columns (copies of G).

Let π be the desired routing permutation we are trying to achieve. Each pebble p has some destination $(\sigma(p), \sigma'(p))$, where $\sigma(p) \in V$, $\sigma'(p) \in V'$. Let us first classify the pebbles according to their second coordinates. Since π is a permutation on $V \times V'$, for each $v' \in V'$, there are exactly $|V|$ pebbles with $\sigma'(p) = v'$. Hence, by the well-known marriage theorem of Hall (see, for example, [11]), we can select a set of distinct representatives from the columns, i.e., one pebble from each column so that their second coordinates are all distinct. Furthermore, we can now repeat this procedure

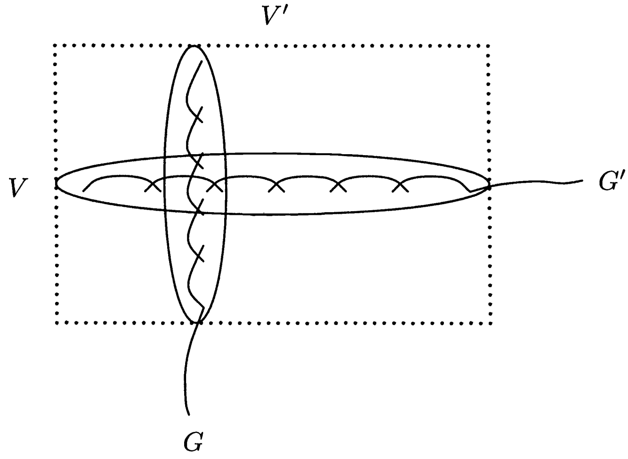


FIG. 4. $G \times G'$.

(again by Hall's theorem) to obtain another set of distinct representatives, and so on. At the end, we see that we can in fact arrange the pebbles in each column so that the pebbles in each row of the rearranged columns all have distinct values of σ'. By hypothesis, this rearrangement can be accomplished in at most $rt(G)$ (parallel) steps.

Next, we rearrange the pebbles in each row (i.e., copy of G') so that pebbles p in the column indexed by $v' \in V'$ have $\sigma'(p) = v'$. This can be done (by hypothesis) in $rt(G')$ more steps, and guarantees, when completed, that the pebbles in each column have distinct values of σ.

The final step, permuting each column (copy of G), can be done in $rt(G)$ more steps. Thus, the whole process requires at most $2rt(G) + rt(G')$ steps.

COROLLARY 1. *For the n-cube Q^n,*

$$rt(Q^n) \leq 2n - 1.$$

COROLLARY 2. *For the m by n grid graph $P_m \times P_n$, $m \leq n$,*

$$rt(P_m \times P_n) \leq 2m + n.$$

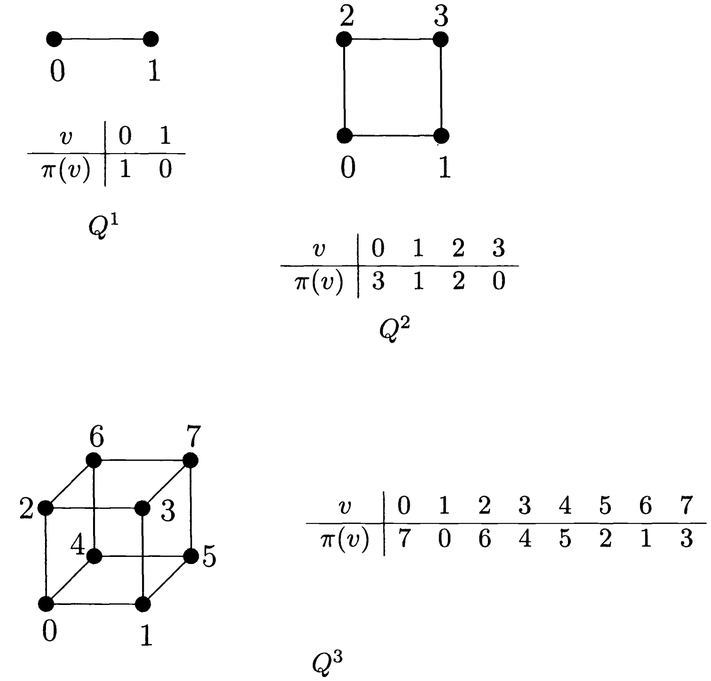
Remarks. Routing on the n-cube Q^n is a very natural question in view of the popular use the n-cube structure for models of parallel computation and communication. Indeed, it was this context (through the work of Ramras [10]) that first motivated our considerations of these questions.

Corollary 1 is well known in the literature. The exact value of $rt(Q^n)$ is still unknown. It is easy to see that $rt(Q^n) \geq n$, since $\text{diam}(Q^n) = n$. The permutations shown in Fig. 5 can be checked to show that $rt(Q^n) \geq n + 1$ for $n = 2, 3$. It is reasonable to conjecture that we always have $rt(Q^n) \geq n + 1$ for $n \geq 2$. Certainly $rt(Q^n) \sim \alpha n$ for some $\alpha \in [1, 2]$. Again, we suspect that the correct value of α is closer to 1 than to 2, but this seems difficult to prove.

6. Flow problems on graphs. Ordinarily, one might expect that $rt(G \times G)$ is substantially larger than $rt(G)$, e.g., as large as $2rt(G)$. However, this is not always the case, as the following result shows.

Let G_n denote the graph consisting of two copies of K_n joined by an edge e (see Fig. 6). It is easy to see that

$$rt(G_n) = 2n + O(1).$$



v	0	1
$\pi(v)$	1	0

Q^1

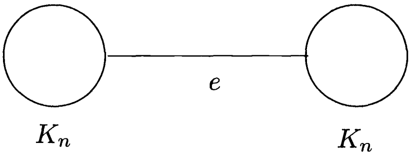
v	0	1	2	3
$\pi(v)$	3	1	2	0

Q^2

v	0	1	2	3	4	5	6	7
$\pi(v)$	7	0	6	4	5	2	1	3

Q^3

FIG. 5. *Bad permutations on Q^n, $n \leq 3$.*



G_n

FIG. 6.

It turns out that $rt(G_n \times G_n)$ is not much larger.

THEOREM 5. *We have*

$$rt(G_n \times G_n) = (1 + o(1))2n. \tag{11}$$

Proof. We can view $G_n \times G_n$ as consisting of 4 copies of $K_n \times K_n$ joined to each other by n parallel edges to form a 4-cycle (see Fig. 7). Within each $V_i = K_n \times K_n$, the two sets of n vertices incident to "crossing" edges have exactly one common vertex. Initially, each V_i has pebbles with destinations lying in the other various V_j's. We group each of V_i's pebbles into sets S_{ij} of size n, according to their destinations (so that all pebbles in a set S_{ij} have the same V_j destination; there may be mixed groups of n left over). We move the S_{ij}'s as a unit, in that all pebbles in each S_{ij} cross from one V_k to another V_l at the same time. Further, we restrict our routing algorithm so that *crossing* moves are only made at *even* times. Of course, permutations *within* a V_i can occur at all times. Since $rt(K_n \times K_n) \leq 6$ (by Theorems 2 and 4), it is not difficult to see that if we have m S_{ij}'s in V_k that should cross over to V_l (which is adjacent to it), then this can be done in $2m + O(1)$ steps. Thus, our problem can be reduced

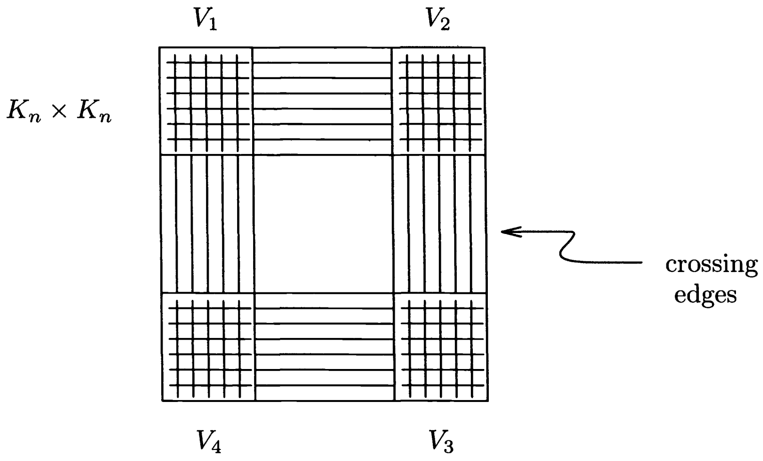


FIG. 7. $G_n \times G_n$.

to the following continuous flow problem on the 4-cycle C_4. We are given one unit of "mass" on each vertex v of C_4 (where we have rescaled our sets of n^2 pebbles at each vertex of C_4 to have total mass 1). Thus, mass is required to "flow" along the edges of C_4 in order to satisfy a 4×4 doubly stochastic *circulation matrix* $C = (C(u,v))$ where, for vertices u, v of C_4, $C(u,v)$ denotes the amount of mass initially at u that must end up at v. Since C is assumed to be doubly stochastic, then $C(u,v) \geq 0$ and

$$\sum_v C(u,v) = 1 = \sum_u C(u,v).$$

Therefore, each vertex of C_4 also ends up with a total of one unit of mass (hence, our use of the terminology "circulation").

In general, a C-circulation φ on $G = (V, E)$ is a set of assignments $\varphi_{uv} : E \to \mathbb{R}^+$, $u, v \in V$, such that, for all u, v,

$$\sum_{ux \in E} \varphi_{uv}(ux) = C(u,v) = \sum_{yv \in E} \varphi_{uv}(yv),$$

while for any $w \neq u, v$,

$$\sum_{sw \in E} \varphi_{uv}(sw) = \sum_{wt \in E} \varphi_{uv}(wt).$$

Intuitively, these equations specify that for each pair $u, v \in V$, $C(u,v)$ units of mass flow from u to v. The norm of φ, denoted by $\|\varphi\|$, is defined to be the maximum amount of mass

$$\varphi(e) = \sum_{u,v} \varphi_{uv}(e)$$

assigned to any edge e of E, where we distinguish between $e = ij \in E$ and the edge $-e = ji$ with the reverse orientation. We say that φ is *balanced* if $\varphi(e) = \varphi(-e)$ for all edges of G.

A little reflection shows that we have proved Theorem 5 if we establish the following lemma.

LEMMA 1. *For all circulation matrices C on C_4, there always exists some balanced C-circulation φ with $\|\varphi\| \le 1$.*

Proof of Lemma 1. Consider a spanning tree T on C_4. There are four such trees; these are all paths of length 3. (See Fig. 8.) Note that there is a unique balanced F-flow φ_T on T. The amount $\varphi_T(e)$ that φ_T assigns to e is just

$$\varphi_T(e) = \sum_{\substack{u \in A \\ v \in B}} C(u,v),$$

where A and B are the components of T formed by the removal of $e = ij$, and $i \in A$, $j \in B$. We form our desired F-flow φ as a convex combination of φ_T's, as T ranges over the spanning subtrees of G. This guarantees that φ is a C-circulation and is balanced (since each φ_T is). In fact, we take the simplest possible convex combination, namely,

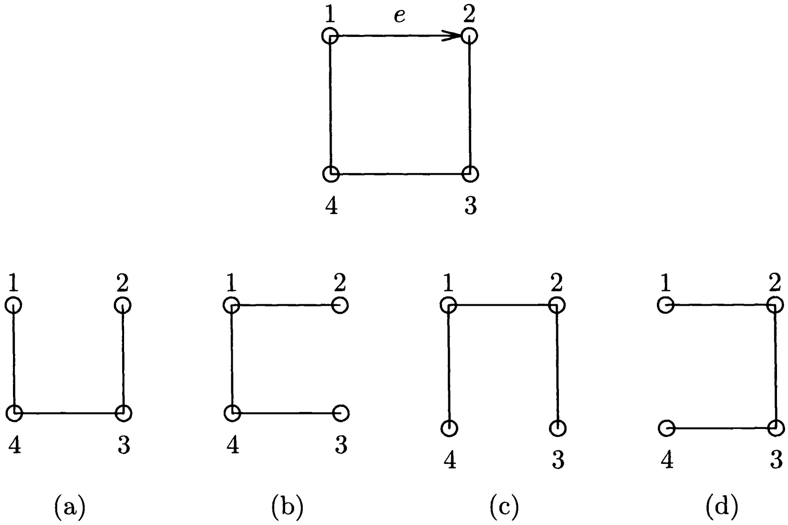


FIG. 8.

$$\varphi := \frac{1}{4} \sum_T \varphi_T,$$

where T ranges over all four spanning trees of C_4. To compute $\|\varphi\|$, we need to bound the value of $\varphi(e)$ for each edge e. For any edge e (by symmetry), the mass $\varphi(e)$ assigned to e is equal to that assigned to the edge in the figure, which is the following:

$$\begin{aligned}
\varphi(e) = 0 & \qquad \text{from (a)},\\
+ \frac{1}{4}(f(1,2) + f(4,2) + f(3,2)) & \qquad \text{from (b)},\\
+ \frac{1}{4}(f(4,2) + f(4,3) + f(1,2) + f(1,3)) & \qquad \text{from (c)},\\
+ \frac{1}{4}(f(1,2) + f(1,3) + f(1,4)) & \qquad \text{from (d)}.
\end{aligned}$$

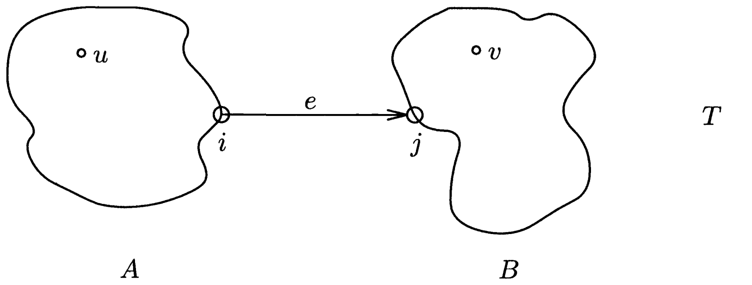


FIG. 9. *Spanning trees of C_4.*

However, observe that in Fig. 9, the maximum value that can flow from A to B is just $\min(|A|, |B|)$. Consequently,

$$f(1,2) + f(4,2) + f(3,2) \le 1,$$
$$f(4,2) + f(4,3) + f(1,2) + f(1,3) \le 2,$$
$$f(1,2) + f(1,3) + f(1,4) \le 1,$$

and so,

$$\varphi(e) \le 1.$$

Since e was arbitrary, $\|\varphi\| \le 1$ and the lemma is proved. □

Now, because φ is balanced, we can reinterpret it as pebble movements, where small time delays (due to the nonuniform S_{ij}, or bounded movement within the V_i's) are negligible as $n \to \infty$. We can then conclude that

$$rt(G_n \times G_n) = (1 + o(1))2n,$$

as claimed. □

The same argument applies, with the same conclusion, for the k-fold product

$$G_n^k = \overbrace{G_n \times \cdots \times G_n}^{k},$$

provided we prove the corresponding flow result on Q^k, the k-cube, which actually is of interest in its own right. We do this with the following theorem.

THEOREM 6. *Let F be a doubly stochastic circulation matrix on Q^k. Then there always exists a balanced C-circulation φ on Q^k with $\|\varphi\| \le 1$.*

Proof. We follow the same strategy as in the case of $C_4 = Q^2$, and build φ as a (uniform) convex combination of tree flows. Define the spanning tree T_k on Q^k recursively as follows:

(i) T_1 is just an edge, which is all of Q^1;

(ii) T_k is formed by adding the edge $e_k = \{(00\ldots0), (10\ldots0)\}$ to join the two copies of T_{k-1} in the corresponding two copies of Q^{k-1} that make up Q^k, namely, $\{\bar{x} = (x_1, \ldots, x_k) \mid x_1 = 0\}$ and $\{\bar{x} = (x_1, \ldots, x_k) \in Q^k \mid x_1 = 1\}$.

We observe that for any edge e of T_k, we can always find a minimum-sized component $A(e)$ of $T_k - \{e\}$ that *does not contain the origin* $(00\ldots0)$. Hence we conclude (by induction) that

$$
\begin{aligned}
g(k) :&= \sum_{e \in T} |A(e)| \\
&= \sum_{e=e_k} |A(e)| + \sum_{e \neq e_k} |A(e)| \\
&= 2^{k-1} + 2 \cdot g(k-1) \\
&= 2^{k-1} + 2 \cdot (k-1) \cdot 2^{k-2} \;=\; k \cdot 2^{k-1},
\end{aligned} \tag{12}
$$

where we check that $g(1) = 1$ satisfies (12) to start the induction.

Next, we construct the family of trees that is used to form φ. Let $\mathrm{Aut}(Q^k)$ denote the automorphism group of Q^k. It is easy to see that $|\mathrm{Aut}(Q^k)| = 2^k \cdot k!$ and that for any $h \in \mathrm{Aut}(Q^k)$, $h(T_k)$ is also a spanning tree of Q^k. Furthermore, for each edge e of Q^k and each edge e' of T_k, there are exactly $2(k-1)!$ choices of $h \in \mathrm{Aut}(Q^k)$ that map e onto e' (as undirected edges; this accounts for the factor of 2. Of course, this does not depend on e' being in T_k.) Define

$$\varphi = \frac{1}{2^k k!} \sum_{T} \varphi_T,$$

where T ranges over all $2^k \cdot k!$ trees $h(T_k)$, $h \in \mathrm{Aut}(Q^k)$. Note that this is just what we did for C_4. To bound $\|\varphi\|$, we compute for any e,

$$\varphi(e) \leq \frac{1}{2^k \cdot k!} \sum_{h \in Aut(Q^k)} \min(|A_h(e)|, |B_h(e)|) = \frac{1}{2^k k!} \cdot 2(k-1)! g(k) = 1.$$

This completes the proof of Theorem 6. □

COROLLARY 3. *For fixed k, if G^k denotes the k-fold Cartesian product $G \times \cdots \times G$ of the graph G shown in Fig.* 6, *then*

$$rt(G^k) = (1 + o(1))2n.$$

Let us define $\mathrm{circ}(G)$, the *circulation index* of a (connected) graph G, by

$$\mathrm{circ}(G) := \sup_{C} \inf_{\varphi} \|\varphi\|, \tag{13}$$

where φ ranges over all balanced C-circulations on G, and C ranges over all doubly stochastic circulation matrices C for G. A trivial lower bound for $\mathrm{circ}(G)$ is the *resistance* of G, defined by

$$\mathrm{res}(G) := \max_{X} \frac{1}{|X|} \min(|A(X)|, |B(X)|), \tag{14}$$

where X ranges over all *cutsets* of G (i.e., minimal sets of edges whose removal disconnects G), and $A(X)$ and $B(X)$ are the connected components formed by removing X. The inequality

$$\mathrm{circ}(G) \geq \mathrm{res}(G) \tag{15}$$

follows by considering the circulation matrix that sends all $|A(X)|$ units of mass into $B(X)$, for an extremal cutset X, where we assume $|A(X)| \leq |B(X)|$. It is interesting to note that (15) holds with *equality* for Q^k. This is not true in general, as can be shown by considering bounded degree expander graphs G^* on n vertices. In that case, we can have

$$\operatorname{res}(G) = O(1) \quad \text{and} \quad \operatorname{circ}(G) > c \log n.$$

Other classes of graphs G for which equality holds in (15) prove interesting. The technique above can be easily used to show that the set of even cycles is such a class.

It seems likely that the space of balanced C-circulations on any graph is spanned by (convex combinations of) the *tree* circulations on G, i.e., the C-circulations in the spanning trees of G.

7. Eigenvalues, random walks and routing. Here we consider d-regular graphs for which all the eigenvalues of the adjacency matrix besides the trivial one have a small absolute value. Let us call a graph G an (n, d, λ)-graph, if it is a d-regular graph on n vertices and the absolute value of every eigenvalue of its adjacency matrix besides the trivial one is at most λ. If λ is small with respect to d, then a random walk on such a graph starting from any vertex converges quickly to the uniform distribution on its vertices. We use this property to derive the following theorem, the proof of which is given later.

THEOREM 7. *Let $G = (V, E)$ be an (n, d, λ)-graph and let σ denote a permutation. Then*

$$rt(G, \sigma) \leq O\left(\frac{d^2}{(d-\lambda)^2} \log^2 n\right).$$

Note that in §2, it is shown that $rt(G)$ is lower bounded by the diameter of G and therefore the routing number of a d-regular graph is at least $\log n / \log(d-1)$ and at most

$$O\left(\frac{d^2}{(d-\lambda)^2} \log^2 n\right).$$

Now define the *expansion coefficient* α of G to be the minimum, over all subsets X of at most half the vertices of G, of the ratio $|N(X) - X|/|X|$, where $N(X)$ is the set of all neighbors of X in G. From §2, we know that the routing number is bounded below by $2/\alpha$. As an immediate corollary of Theorem 7, up to a polylogarithmic factor, $rt(G)$ is bounded above by a polynomial in $1/\alpha$ for any regular graph with polylogarithmic degrees.

COROLLARY 4. *If $G = (V, E)$ is a d-regular graph on n vertices with expansion coefficient α, then*

$$rt(G) \leq O\left(\frac{d^2}{\alpha^4} \log^2 n\right).$$

Proof. The main result of [1] states that if α is the expansion coefficient of a d-regular graph, then the second largest eigenvalue of its adjacency matrix is at most $d - \frac{\alpha^2}{4+2\alpha^2}$. Suppose, first, that this is an upper bound for the absolute value of every negative eigenvalue as well. Then, by Theorem 7, $rt(G) = O(l^2)$ for $l = O(\frac{d}{\alpha^2} \log n)$. If there are negative eigenvalues of large absolute value, we first add d loops in every vertex and apply the result to the new graph. This completes the proof. □

In a similar way, we define the *edge expansion coefficient* β of G to be the minimum, over all subsets X of at most half the vertices of G, of the ratio $|\Gamma(X)|/|X|$, where $\Gamma(X)$ is the set of edges of G leaving X (i.e., with exactly one endpoint in X). We remark that the inverse of the edge expansion coefficient is exactly the resistence of G. From §2, we know that the routing number is bounded below by $2/\beta - 1$. As a corollary of Theorem 7, up to a polylogarithmic factor, $rt(G)$ is bounded above by a polynomial in $1/\beta$ for any regular graph with polylogarithmic degrees.

COROLLARY 5. *If $G = (V, E)$ is a d-regular graph on n vertices with edge expansion coefficient β, then*

$$rt(G) \leq O\left(\frac{d^4}{\beta^4}\log^2 n\right).$$

Proof. The proof follows from the well-known fact (see [8]) that if β is the edge expansion coefficient of a d-regular graph, then the second largest eigenvalue of its adjacency matrix is at most $d - \frac{\beta^2}{2d}$. □

Note that by the above two corollaries, $rt(G) \leq O(\log^2 n)$ for any bounded degree expander on n vertices (i.e., any regular bounded degree graph on n vertices with expansion coefficient or edge expansion coefficient bounded away from 0).

Many interconnection networks studied in the literature are, in fact, Cayley graphs. A simple corollary of the above theorem implies that the routing number of a Cayley graph is intimately related to its diameter.

COROLLARY 6. *For any Cayley graph G of a group of n elements with a polylogarithmic (in n) number of generators, the diameter of G is polylogarithmic, if and only if the routing number $rt(G)$ is polylogarithmic.*

Proof. As shown in [3], a Cayley graph of polylogarithmic diameter has an inverse polylogarithmic expansion coefficient; hence the result follows from Corollary 4. □

The proof of Theorem 7 follows from the following lemmas. Our first lemma holds for any d-regular graph G. A *random walk* of length l starting at a vertex v of G is a randomly chosen sequence $v = v_0, v_1, \ldots, v_l$, where each v_{i+1} is chosen, randomly and independently, among the neighbors of v_i, $(0 \leq i < l)$. We say that the walk *visits v_i at time i*. We make no attempt to optimize the constants here and in what follows.

LEMMA 2. *Let $G = (V, E)$ be a d-regular graph on n vertices, and suppose that $l \geq \log n$. For any $v \in V$ independently, let $P(v)$ denote a random walk of length l starting at v. Let $I(v)$ denote the total number of other walks $P(u)$, such that there exists a vertex x and two indices $0 \leq i, j \leq l$, $|i - j| < 5$ so that $P(v)$ visits x at time i and $P(u)$ visits x at time j. Then, almost surely (i.e., with probability that tends to 1 as n tends to infinity), there is no vertex v such that $I(v) > 100(l + 1)$.*

Proof. Let A be the normalized adjacency matrix of G, i.e., the matrix $A = (a_{uv})_{u,v \in V}$ defined by $a_{uv} = l(u, v)/d$, where $l(u, v)$ is the number of edges between u and v. The probability that the random walk $P(u)$ visits x at time i is precisely $e(x)^t A^i e(u)$, where $e(y)$ is the unit vector having 1 in coordinate y and 0 in any other coordinate. Given the random walk $P(v)$ and a value of i, $0 \leq i \leq l$, there is a unique vertex $x = x(v, i)$ in which $P(v)$ visits at time i. For any given $u \neq v$, the conditional probability that for some j satisfying $|i - j| < 5$, the walk $P(u)$ visits x at time j is thus at most $e(x)^t \sum_{j:|j-i|<5} A^j e(u)$. The probability $p(v, u)$ that there exists some vertex x and two indices $0 \leq i, j \leq l$, $|i - j| < 5$ follows, so that $P(v)$ visits x at time i and $P(u)$ visits x at time j can be bounded by

$$p(v, u) \leq \sum_{i=0}^{l} e(x(v, i))^t \sum_{j:|j-i|<5} A^j e(u).$$

By summing over all possible starting points u (including v itself, where this last summand corresponds to adding another independent random walk starting at v, an addition which may only increase the expectation of $I(v)$), we conclude that the expectation of $I(v)$ is at most

$$\sum_{u \in V} p(v,u) \leq \sum_{i=0}^{l} e(x(v,i))^t \sum_{j:|j-i|<5} A^j e,$$

where e is the all 1 vector. Since e is an eigenvector of A with eigenvalue 1, the last expression can be computed precisely by showing that it is strictly less than $10(l+1)$. We have thus shown that for each fixed v, the expectation of the random variable $I(v)$ is strictly less than $10(l+1)$. Observe that this random variable is a sum of $n-1$ independent indicator random variables whose expectations are the quantities $p(v,u)$. It thus follows easily from the known estimates for large deviations of sums of independent indicator random variables (see, e.g., [2, Thm. A.12, p. 237]), that for each fixed v, the probability that $I(v)$ exceeds, say, $100(l+1)$ is at most

$$(e^9/10^{10})^{10(l+1)} << 1/n^2.$$

(A similar estimate can in fact be proved directly. Given a set of m independent events, with the probability of the ith event being p_i, suppose that $\sum p_i \leq r$. Then, the probability that at least s events occur can be bounded by

$$\sum_{S \subset \{1,\ldots,m\},|S|=s} \Pi_{i \in S}\ p_i \leq \frac{1}{s!}\left(\sum p_i\right)^s \leq (re/s)^s.$$

In our case, we have $r = 10(l+1)$ and $s = 10r$.)

Since there are only n vertices v, the probability that there is a vertex v with $I(v) > 100(l+1)$ is (much) smaller than $1/n$, completing the proof. □

LEMMA 3. *Let $G = (V,E)$ be an $(n,d,(1-\epsilon)d)$-graph and let σ be a permutation of order two of V (i.e., a product of pairwise disjoint transpositions). Put $l = \frac{10}{\epsilon}\log n$. Then there is a set of $n/2$ walks $P(v) = P(\sigma(v))$, $v \in V$, of length $2l$ each, where $P(v)$ connects v and $\sigma(v)$ such that the following holds. Let $I(v)$ denote the total number of other walks $P(u)$ such that there exists a vertex x and two indices $0 \leq i,j \leq l$, $|i-j| < 5$, so that $P(v)$ visits x at time i, and $P(u)$ visits x at time j or at time $2l-j$. Then $I(v) \leq 400(l+1)$ for all v.*

Proof. Let $P(v)$ be a random walk of length $2l$ between v and $\sigma(v)$. As shown in [6] (using an argument similar to the one used previously in [12]), we may assume that each walk $P(v)$ consists of two random walks of length l each, one starting from v and one from $\sigma(v)$. The reason for this is that by our eigenvalue condition, a random walk of length l is almost uniformly distributed on the vertices of G, and hence one may view the walk $P(v)$ as being chosen by first choosing its middle point (according to a uniform distribution) and then by choosing its two halves. For more details, see [6]. The result thus follows from Lemma 2. □

Proof of Theorem 7. Let $G = (V,E)$ be an (n,d,λ)-graph. It suffices to consider a permutation σ of order two of V (i.e., a product of pairwise disjoint transpositions), since any permutation is a product of at most two such permutations (as proved in Theorem 2). We set $\epsilon = 1-\frac{\lambda}{d}$ and $l = \frac{10}{\epsilon}\log n$. We want to show that $rt(G,\sigma) < O(l^2)$. Let $P(v)$ be a system of walks of length $2l$ satisfying the assumption of the previous corollary. Let H be the graph whose vertices are the walks $P(v)$ in which $P(u)$ and $P(v)$ are adjacent, if there exists a vertex x and two indices $0 \leq i,j \leq l$, $|i-j| < 5$, so that $P(v)$ visits x at time i, and $P(u)$ visits x at time j or at time $2l-j$. Then the maximum degree of H is $O(l)$ and hence it is $O(l)$-colorable. It follows that we can split all our paths $P(v)$ into $O(l)$ classes of paths such that the paths in each class are not adjacent in H. Consider now the following routing algorithm. For each set of paths as above, perform $2l+1$ steps, where the step numbers i and $2l+2-i$ correspond to flipping the pebbles along edge numbers i and $2l+1-i$ in each of the paths in the set for all $i<l$. Step number l flips edge l, and step $l+1$ flips edge $l+1$. We can check that by the end of these $2l+1$ steps, the ends of each path exchange pebbles, and all the other pebbles stay in their original places. (Note that some pebbles that are not at the ends of any of the paths may move several times during these steps, but the symmetric way these are performed guarantees that such pebbles will return to their original places at the completion of the $2l+1$ steps.) By repeating the above for all the path-classes, the result follows. □

8. The route covering number of a graph. We next discuss several problems closely related to the routing number of a graph. One such problem is the following.

Suppose $G=(V,E)$ is a connected graph on n vertices. For a permutation π, we consider a *route set* P, which is just some set of paths P_i joining each vertex v_i to its destination vertex $\pi(v_i)$, for $i=1,\ldots,n$. For each edge e of G, we consider the number $\mathrm{rc}(e,G,\pi,P)$ of paths P_i in P which contain e. The *route covering number* $\mathrm{rc}(G)$ of G is defined to be

$$\mathrm{rc}(G)=\max_{\pi}\ \min_{P}\ \max_{e\in E}\ \mathrm{rc}(e,G,\pi,P).$$

In other words, for each permutation, we want to choose the route set so that the maximum number of occurrences of any edge in the paths of the route set is minimized. It is easy to see that the route covering problem is a special case of C-circulation obtained by choosing C to satisfy $C(u,v)=1$ if $v=\pi(u)$, and 0 otherwise, for each permutation π, and by insisting on integer valued circulation.

For example, for the n-cube Q^n, the method described in Theorem 4 gives

$$\mathrm{rc}(Q^n)\le 4.$$

In the other direction, by choosing π to be the permutation of vertices in Q^n so that the distance between v and $\pi(v)$ is n for every vertex v, it can easily be seen that

$$\mathrm{rc}(Q^n)\ge\frac{\sum_v \mathrm{dist}(v,\pi(v))}{|E(Q^n)|}=2.$$

The problem of determining the exact value of $\mathrm{rc}(Q^n)$ for general n remains unresolved. Also of interest is a "symmetric" version of the route covering problem especially for Q^n.

An *assignment* for Q^n is a partition of the vertex set of Q^n into subsets of size 2 or less. Is it possible to find edge-disjoint paths joining vertices in the same subset for any assignment of Q^n ?

The answer is negative when n is even. However, for odd n this problem remains open.

9. Concluding remarks. Numerous unanswered questions remain, some of which we now mention.

(1) Is it true that, for any tree T_n on n vertices,

$$rt(T_n)\le\frac{3}{2}n+o(n)?\qquad \frac{3}{2}n+O(1)?$$

(2) Is it true that, for the n-cube Q^n,

$$rt(Q^n) = n + o(n)? \qquad n + O(1)?$$

(3) Is it true that, for every graph G,

$$rt(G \times G) \geq rt(G)?$$

(4) Is it true that, for an expander graph G of bounded degree,

$$rt(G) = O(\log n)?$$

(5) Characterize graphs G with $\operatorname{circ}(G) = \operatorname{res}(G)$.

(6) Are the balanced C-circulations on a graph always spanned by the spanning tree C-circulations on the graph?

(7) What is the computational complexity of determining $rt(G)$?

REFERENCES

[1] N. Alon, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83–96.

[2] N. Alon and J. H. Spencer, *The Probabilistic Method*, John Wiley, New York, 1991.

[3] L. Babai and M. Szegedy, *Local expansion of symmetrical graphs*, Combin., Probab. and Comput., 1 (1991), pp. 1–12.

[4] M. Baumslag and F. Annexstein, *A unified framework for off-line permutation routing in parallel networks*, Math. Systems Theory, 24 (1991), pp. 233–251.

[5] V. E. Beněs, *Mathematical Theory of Connecting Networks*, Academic Press, New York, 1965.

[6] A. Broder, A. Frieze, and E. Upfal, *Existence and construction of edge disjoint paths on expander graphs*, in Proc. 24th ACM Sympos. on Theory of Comput., Victoria, British Columbia, Canada, May 1992, pp. 140–149.

[7] W. Goddard, private communication.

[8] M. Jerrum and A. Sinclair, *Approximating the permanent*, SIAM J. Comput., 18 (1989), pp. 1149–1178.

[9] D. E. Knuth, *The Art of Computer Programming*,Vol. 3, Addison–Wesley, Reading, MA, 1973, p. 241.

[10] M. Ramras, *Routing permutations on a graph*, Networks, 23 (1993), pp. 391–398.

[11] H. J. Ryser, *Combinatorial Mathematics*, Carus Monograph No. 14, Mathematical Association of America, John Wiley, New York, 1963.

[12] L. Valiant, *A scheme for fast parallel communication*, SIAM J. Comput., 11 (1982), pp. 350–361.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 531–543, November 1994

PREDICTING CAUSE-EFFECT RELATIONSHIPS FROM INCOMPLETE DISCRETE OBSERVATIONS *

E. BOROS[†§], P. L. HAMMER[†§], AND J. N. HOOKER[‡]

Abstract. This paper addresses a prediction problem occurring frequently in practice. The problem consists in predicting the value of a function on the basis of discrete observational data that are incomplete in two senses. Only certain arguments of the function are observed, and the function value is observed only for certain combinations of values of these arguments. The problem is considered under a monotonicity condition that is natural in many applications. Applications to tax auditing, medicine, and real estate valuation are discussed. In particular, a special class of problems is identified for which the best monotone prediction can be found in polynomial time.

Key words. cause-effect relationship, monotone regression, incomplete observations

AMS subject classifications. 62J02, 06A10, 90C09

1. Introduction.

The problem of establishing cause-effect relationship based on incomplete observations was studied in [7]. In this paper we address the problem of finding a good approximation of an unknown discrete function on the basis of a set of observations, which is incomplete in two senses. We observe the values of only some of the arguments of the function and we observe the function value only for certain combinations of values of these arguments. Our goal is to predict the value of the function for any combination of values for these arguments. Such problems occur frequently, and we begin with some examples.

Suppose that a tax bureau must decide which tax forms to audit, with the goal of auditing only those for which the increased return justifies the auditing expense. The observation set consists of data records corresponding to forms audited in the past. Each record indicates whether the audit was justified and lists some attributes of the taxpayer. The problem is to determine, on the basis of these attributes, when future tax forms should be audited.

A taxpayer's attributes form a vector x of Boolean (0-1) variables, where each x_j indicates whether a certain threshold is exceeded. For example, one x_j may have the value 1 when the taxpayer claims too many dependents for a person of his age, and another may have the value 1 when he claims too many charitable contributions relative to his income.

The observation set partially defines a function g whose value is 1 when auditing was justified, and 0 otherwise. Whether an audit was justified depends not only on the recorded attributes used to make the auditing decision, but on a number of hidden factors as well. The arguments of g therefore consist of a vector x of the recorded

* Received by the editors April 16, 1991; accepted for publication (in revised form) July 16, 1993.

[†] Rutgers Center for Operations Research (RUTCOR), Rutgers University, P.O. Box 5062, New Brunswick, New Jersey 08903.

[‡] Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. This author's research was supported by Air Force Office of Scientific Research grant 91-0287 and Office of Naval Research grant N00014-92-J-1028.

[§] This author's research was partially supported by Air Force Office of Scientific Research grants 89-0512B and F49620-93-1-0041 and Office of Naval Research grants N00014-92-J-1375 and N00014-92-J-4083.

attributes and a vector y of hidden attributes. So, each data record can be written in the form $(x, g(x, y))$. In practice, the function g is only partially defined because its value $g(x, y)$ is not given for every possible vector (x, y) of attributes. We cannot even assume there is an observation for every possible x.

Our problem is to predict the value of $g(x, y)$ on the basis of x alone. To do this, we derive an *approximation* f of g that is a function of x only. Deriving f requires that we interpolate a value when x has not been observed. However, even when x has been observed, we may need to reconcile different observed values of $g(x, y)$ for different y's.

We focus on problems in which f is *monotone*, in the sense that $f(x) \not\succ f(x')$ whenever $x \preceq x'$, where $\prec$ is a partial order on the outcomes and $\prec$ a partial order on the x's. In the tax auditing example, the two outcomes are linearly ordered $(0 \prec 1)$, and $\prec$ is defined by $x \preceq x'$ whenever $x_j \leq x'_j$ for all j. Monotonicity is a reasonable assumption for this problem, because every 1 among the attributes is another reason for suspecting that the form should be audited. If one form with certain suspicious traits should be audited, then another form with these and still other suspicious traits should certainly be audited.

We also generalize our approach to problems in which more than two outcomes are possible, where these outcomes are partially ordered. The tax auditor, for example, might classify forms as needing no audit, needing a second look to determine the desirability of an audit, and needing an audit right away. These outcomes happen to be linearly ordered in a natural way (in the order listed), and it is reasonable to assume that f is monotone.

The attributes as well as the outcome can have more than two values. Suppose, for example, that we have a battery of 45 biochemical tests for carcinogenicity, each of which can have outcomes *negative*, *indefinite*, and *positive*. We do not apply every test to a given chemical; when a test is not applied, we say that the test "result" is *no data*. The 45 test results are viewed as attributes of the chemical tested. Since the tests are not foolproof, we first apply them to a number of chemicals that have been previously tested clinically for carcinogenicity, with possible outcomes *harmless*, *undetermined* (indicating inconclusive clinical experience), *potentially dangerous* (meaning that the chemical causes cancer when ingested in very large doses), and *dangerous*. The problem is to find a function f that predicts the outcome of clinical trials for a new chemical.

To check for monotonicity, we must impose a partial ordering on the attribute vectors and on the outcomes. The four possible values of each attribute x_j submit to a partial order: *negative* is less than *indefinite* and *no data*, which are less than *positive*; but *indefinite* and *no data* are incomparable. We can therefore say that $x \preceq x'$ when $x_j \preceq x'_j$ for all j. The four outcomes have a similar partial ordering: *harmless* is less than *potentially dangerous* and *undetermined*, both of which are less than *dangerous*; but we may be unable to order *potentially dangerous* and *undetermined* with respect to each other. It is reasonable to assume that f is monotone with respect to these orderings.

Whenever f incorrectly predicts the outcome for a given observation, a penalty is incurred, where the size of the penalty depends on the correct and predicted outcomes. For us, the best approximation f is one that minimizes the total penalty. In [4] we investigate the issue of defining "best approximation" from a statistical point of view and distinguish the analysis described here from logit and categorical data analysis.

If there is no restriction on f, finding the best approximation is relatively easy, since we can treat each x separately. That is, for any fixed x, $f(x)$ should have the value that minimizes the sum of the penalties over observations of the form (x, y). If x occurs in no observations, $f(x)$ can be set to an arbitrary value. However, if f is required to be monotone (or to have some other restrictive property), the problem cannot in general be decomposed this way, and the observed values of $f(x)$ for unobserved x are in general restricted. In [4] we consider some other possible restrictions on f (when it is a Boolean function). More examples for similar problems can be found in [6], [21], [22].

In this paper, we show that a network flow model can be used to determine the best approximation f when the partial ordering of the outcomes belongs to a special class of partial orders, including interval orders. An *interval order* is one in which every element can be associated with an interval of real numbers, such that $\alpha \prec \beta$ if and only if the upper end of the interval associated with α is smaller than the lower end of that associated with β. A special case of an interval order is one that is layered, in the sense that the elements are partitioned into a sequence of sets $S_1, \ldots, S_m$, such that all the elements in each S_i are incomparable, but everything in S_i is less than or equal to everything in S_{i+1}. The outcomes of the clinical trials for carcinogenicity have this sort of ordering. A linear order, as in the tax auditing problem, is, of course, an interval order.

Another instance that might call for an interval ordering is estimating the value of a piece of property. Suppose that we suspect that certain combinations of property attributes justify assessing the value to be within certain ranges. The ranges may overlap, but the outcomes nonetheless have an interval order. If higher-ranked attribute values are more desirable, it may be reasonable to assume monotonicity.

We begin in the next section with a precise statement of the problem and a small example. In §3 we show that finding the best monotone approximation is equivalent to a generalization of the "maximal closure" problem on a directed graph [16]. When there are two ordered outcomes, the generalization coincides with the maximal closure problem. In §4 we describe a special class of partial orders, the so-called aligned orders. We prove that aligned orders can be recognized in polynomial time and that interval orders are aligned. Finally, in §5 we show that the generalized maximal closure problem can be reduced to the maximal closure problem when the outcomes form an aligned order. Since the maximal closure problem can be solved via a minimum cut computation, we can solve the best approximation problem likewise. The complexity of our algorithm is $O(n^3m^3)$, where n is the number of distinct input attribute vectors x and where m is the number of possible outcomes.

2. Problem statement and main results. Let $(X, \prec)$ be a finite partially ordered set of (observed) attribute vectors x and $(V, \prec)$ a finite partially ordered set of outcomes. The mapping $\mu : X \times V \longmapsto \mathbb{Z}$ indicates the number of times an attribute vector x resulted in a given outcome; that is, $\mu(x, \alpha)$ is the number of observations (x, y) for which $g(x, y) = \alpha$. The penalty function is $c : V \times V \longmapsto \mathbb{Z}$, where $c_{\beta,\alpha}$ is the penalty for predicting outcome β when the observed outcome is α. The problem is to find a *V-monotone* function $f(x)$, i.e., one for which $f(x) \not\succ f(x')$ whenever $x \preceq x'$, so as to minimize the total penalty,

$$\varepsilon[f] = \sum_{x \in X} \sum_{\alpha \in V} c_{f(x),\alpha} \mu(x, \alpha). \tag{2.1}$$

As a matter of fact, we want to define the function $f(x)$ only on the set X of observed attribute vectors. For an unobserved value x', we can let $f(x')$ take any value for which $f(x') \not\succ f(x)$ for all observed values x satisfying $x' \preceq x$ and for which $f(x) \not\succ$

$f(x')$ for all observed values x satisfying $x \preceq x'$. The above problem is called the *monotone approximation problem*, and an instance of this problem is denoted by $\mathcal{M} = (X, V, c, \mu)$.

Let us observe that the objective function in the above problem is separable in the sense that it can be written as

$$\varepsilon[f] = \sum_{x \in X} h_x(f(x)),$$

where

$$h_x(f(x)) = \sum_{\alpha \in V} c_{f(x),\alpha}\mu(x, \alpha),$$

and the function h_x depends only on x but not on other elements of X. The problem of separable optimization subject to precedence (-like) constraints, in general, can be formulated as follows:

$$\min \sum_{x \in X} h_x(f(x))$$

$$\text{(SOP)} \qquad \begin{aligned} &\text{s.t. } f(x) \in V \quad \text{for all } x \in X, \quad \text{and} \\ &f(x) \not\succ f(y) \quad \text{whenever} \quad x \prec y. \end{aligned}$$

Here $(X, \succ)$ and $(V, \succ)$ are partially ordered sets, X is a finite set, and $h_x : V \to \mathbb{R}$ are given real-valued functions for $x \in X$.

If $(V, \prec)$ is the set of reals with the usual order $<$ and if h_x are convex functions, (SOP) is known in the statistical literature as the *isotonic regression* problem; see, e.g., [2]. There is a bewildering variety of algorithms created to solve this problem in various special cases, many of which are not polynomial. An equivalent problem was considered in [14], [19] in the context of inventory/production systems. An $O(|X|^4)$ algorithm was provided in these papers for the case when the functions h_x are special convex differentiable functions for $x \in X$. For general convex functions and for the case of handling general upper and lower bounds, an algorithm of similar complexity is given in [18]. For a somewhat more general class of objective functions, a different algorithm of the same complexity is presented in [5]. The methods of the last two papers extend to the case when $(V, \succ)$ is a linearly ordered discrete (finite or infinite) set. In the further specialized case, when $(X, \succ)$ is a linear order and h_x are convex quadratic real functions, an $O(|X|)$-time algorithm is provided in [3].

None of the above methods can handle the problem (SOP) in polynomial time if the objective function is not special (usually convex, or something almost as restrictive). The authors are not aware of any results for the case where $(V, \succ)$ is not linearly ordered.

In this paper, we address the problem (SOP) in the case when $(V, \succ)$ is a finite partially ordered set and no restrictions are imposed on the functions $h_x, x \in X$. The main result of this paper is that the problem (SOP) can be solved in polynomial time in the case when $(V, \succ)$ belongs to a special class of partially ordered sets. We say that a partial order $(V, \succ)$ is *aligned* if there is a linear extension of $\succ$, i.e., a labeling $V = \{\nu_0, \dots, \nu_v\}$ of the elements of V satisfying the following two conditions:

TABLE 1
A sample observation set.

			$\mu(x,\alpha)$				$\sum_{\beta} c_{\alpha\beta}\mu(x,\beta)$									
	x_1	x_2	h	u	p	d	h	u	p	d	f_0	f_1	w_{xh}	w_{xu}	w_{xp}	w_{xd}
1	N	N	2	1	0	0	2	4	5	8	h	h	−2	−2	−3	−1
2	I	N	1	1	0	1	5	4	5	5	u	u	−5	1	0	−1
3	I	U	1	0	0	0	0	2	2	3	h	u	0	−2	−2	1
4	P	U	1	1	2	0	6	4	3	9	p	p	−6	2	3	−8
5	I	P	0	0	1	2	8	5	4	2	d	d	−8	3	4	−1

$$\text{(C1)} \qquad \forall \nu_i \prec \nu_j \quad \exists \text{ an index } k \text{ such that } i \leq k,\ k+1 \leq j, \text{ and } \nu_k \prec \nu_{k+1};$$

$$\text{(C2)} \qquad \nu_k \prec \nu_{k+1} \Rightarrow \nu_j \prec \nu_{k+1} \quad \forall j \leq k.$$

We show that (SOP) can be solved in $O(|X|^3|V|^3)$ time if $(V,\succ)$ is aligned. We also show that, given a partial order $(V,\succ)$, we can recognize in $O(|V|^2)$ time whether it is aligned, and we prove that interval orders are aligned.

We use the following as a running example throughout the remainder of the paper. Suppose that we wish to determine the carcinogenicity of chemicals on the basis of only two tests. We have clinical data for five chemicals (Table 1). N, U, I, and P, respectively, denote the possible test results (*negative*, *no data*, *indefinite*, and *positive*), and h, u, p, and d denote the four possible outcomes (*harmless*, *undetermined*, *potentially dangerous*, and *dangerous*). The table displays the number $\mu(x,\alpha)$ of times chemicals with attribute vector x lead to clinical result α.

To define the penalty for error, let us rank the outcomes as follows: h has rank 1, u and p have rank 2, and d has rank 3. Then we might say that the penalty for an error is 1 more than the distance between the ranks of the predicted and observed outcome when the outcomes are different, and 0 otherwise. For instance, predicting a potentially dangerous (p) chemical to be harmless (h) brings penalty 2, and predicting it to be undetermined (u) brings penalty 1. The total penalty $\sum_{\beta} c_{\alpha\beta}\mu(x,\beta)$ for predicting that chemicals described by x have clinical outcome α appears in the middle of the table. For instance, the penalty for predicting that chemicals with test results (N,N) are dangerous is 8, because they were observed to be harmless on two occasions and to have undetermined effect on one occasion.

If there are no restrictions on the approximating function f, the best approximation is clearly the function f_0 shown in the table. We set $f_0(N,N) = h$, for instance, since predicting h results in the least penalty ($\sum_{\beta \in V} c_{h,\beta}\mu((N,N),\beta) = 2c_{h,h} + c_{h,u} = 2$). In general, f_0 is defined such that the sum $\sum_{\beta \in V} c_{f_0(x),\beta}\mu(x,\beta)$ is minimal for $x \in X$. In this example, however, f_0 is not monotone, since $f_0(I,N) \succ f_0(I,U)$ even though $(I,N) \preceq (I,U)$. Our task is to find the best monotone approximation.

3. Maximal V-partitions. Let $G = (N, A, w)$ be a directed graph, where N denotes the set of nodes and where A denotes the set of arcs. A subset $C \subseteq N$ of the nodes is called a *terminal set* of G if $(x,y) \in A, x \in C$ implies that $y \in C$. Clearly, the intersection of two terminal sets is again a terminal set. For a given subset S of the nodes, its *closure* $\mathrm{cl}(S)$ is defined as the minimal terminal set containing S. In

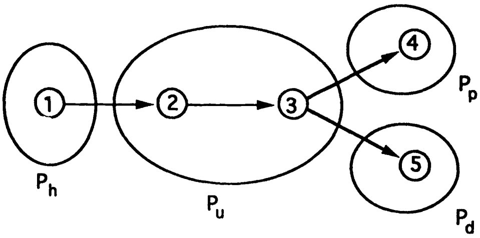


FIG. 1. *Illustration of a V-chain.*

other words, cl(S) contains exactly those vertices y of G that are either belonging to S or accessible from S by a directed path of G. Given real weights $w_x, x \in N$ associated with the nodes, a *maximal closure* of the graph G is a terminal set C for which $w(C) = \sum_{x \in C} w_x$ is maximal.

To an instance $\mathcal{M} = (X, V, \mu, c)$ of the monotone approximation problem, we associate a directed graph $G_{\mathcal{M}} = (X, A)$ in which the observed attribute vectors are the nodes and where there is an arc from x to y whenever $x \prec y$.

As an example, consider the graph G_0 in Fig. 1, which has nodes corresponding to the five values of x in Table 1. (Ignore for now the large sets circled in Fig. 1.) G_0 contains an arc (x, x') whenever $x \preceq x'$. The closure of a set C of nodes (x's) is the set of all x's greater than or equal to the x's in C. For instance, the closure of $\{3\}$ is $\{3, 4, 5\}$.

Picard [16] introduced the problem of finding a maximal closure in a given directed graph and showed that the maximal closure can be found as the source side of a minimum cut in an associated network. This problem has many interesting applications, e.g., [1], [11]–[13], [15], [17], [20], [23].

In this section, we introduce a generalization of this problem, which is shown to be equivalent to the problem of V-monotone approximation and which is reducible in certain cases to the problem of maximal closures.

Let us consider a partially ordered finite set $(V, \prec)$ and let $\alpha || \beta$ denote the fact that α and β are incomparable elements of V. A partition $\mathcal{P} = \{P_\alpha | \alpha \in V\}$ of the nodes of G (i.e., $\bigcup_{\alpha \in V} P_\alpha = N$) is called a *V-partition* if

$$(3.1) \qquad \forall\,(i, j) \in A, \quad i \in P_\alpha \quad \text{and} \quad j \in P_\beta \quad \text{imply that } \alpha \not\succ \beta.$$

For example, suppose that we let $V = \{h, u, p, d\}$, as in Table 1. One possible V-partition is given by the circled sets in Fig. 1, that is, $P_h = \{1\}, P_u = \{2, 3\}, P_p = \{4\}$, and $P_d = \{5\}$.

Note that the sets in this V-partition are determined by the function f_1, displayed in Table 1. Namely, the set P_d is the set of nodes classified by f_1 as dangerous $(f_1(x) = d), P_p$ is the set of nodes classified as potentially dangerous $(f_1(x) = p), P_u$ is the set of nodes classified as undetermined $(f_1(x) = u)$, and P_h is the set of nodes classified as harmless $(f_1(x) = h)$.

Note that the function f_1 that defines the above V-partition of Fig. 1 is V-monotone. In fact, we have the following result.

THEOREM 1. *Let $G_{\mathcal{M}} = (X, A)$ be the directed graph associated with a given instance of the monotone approximation problem. Then there is a one-to-one correspondence between the V-monotone functions and the V-partitions of G.*

Proof. Let us assume first that $f : X \longmapsto V$ is a V-monotone function. For $\alpha \in V$, let

$$(3.2) \qquad P_\alpha = \{x \mid f(x) = \alpha\}$$

and let $\mathcal{P}_f = \{P_\alpha | \alpha \in V\}$. We show that $\mathcal{P}_f$ is a V-partition.

Now condition (3.1) is implied by the monotonicity of f, as follows. Let $(x, y) \in A$ be an arbitrary arc with $x \in P_\alpha$ and with $y \in P_\beta$. We now have $\alpha = f(x)$ and $\beta = f(y)$ by (3.2). The monotonicity of f implies that

$$f(x) \not\succ f(y). \tag{3.3}$$

It follows then that either $\alpha \preceq \beta$ or $\alpha || \beta$, and hence (3.1) follows.

For the converse relation, let $\mathcal{P} = \{P_\alpha | \alpha \in V\}$ be a V-partition of $G_{\mathcal{M}}$. Then the function $f_{\mathcal{P}}$, defined by

$$f_{\mathcal{P}}(x) = \alpha \quad \text{if } x \in P_\alpha, \tag{3.4}$$

is clearly an $X \longmapsto V$ mapping, and property (3.1) immediately implies its V-monotonicity. □

Let $w_{x,\alpha}$ be given real weights for $x \in N$ and $\alpha \in V$ and let the weight $w(\mathcal{P})$ of the V-partition $\mathcal{P}$ be defined as

$$w(\mathcal{P}) = \sum_{\alpha \in V} \sum_{\beta \succeq \alpha} \sum_{x \in P_\beta} w_{x,\alpha}. \tag{3.5}$$

Then a *maximum-weight V-partition* of the weighted graph $G = (N, A)$ is a V-partition $\mathcal{P}$, which maximizes $w(\mathcal{P})$.

Let us again consider the directed graph $G_{\mathcal{M}} = (X, A)$ associated with an instance $\mathcal{M} = (X, V, \mu, c)$ of the monotone approximation problem. Let us now define weights $w_{x,\alpha}$ for $x \in X$ and $\alpha \in V$ by the following set of equations:

$$\sum_{\beta \preceq \alpha} w_{x,\beta} = -h_x(\alpha) \left(= -\sum_{\beta \in V} c_{\alpha,\beta} \mu(x, \beta) \right) \quad \text{for all } \alpha \in V. \tag{3.6}$$

It is easy to observe that, if we arrange the elements of V according to an arbitrary linear extension of $\prec$, then the coefficient matrix on the left-hand side of (3.6) becomes a lower triangular matrix with 1's in the main diagonal. Hence, for any integral parameter vectors c and μ, the system of equations (3.6) has a unique integral solution.

We show below that the problem of monotone approximation is equivalent to a maximum-weight V-partition problem in the associated graph.

We know that finding a V-partition is equivalent to finding a monotone approximation f, but, to equate the maximal V-partition problem to the monotone approximation problem, we must relate the weight of a V-partition to the penalty associated with an approximation. This is done in the example of Table 1 as follows. The negative of the weight w_{xh} contributed by each x lying in P_h to a V-partition is the penalty that results from predicting chemicals with attributes x to be harmless. The negative of the total weight $w_{xh} + w_{xu}$ contributed by each x in P_u is the penalty for predicting chemicals with attributes x to be undetermined, and similarly for $w_{xh} + w_{xp}$. The negative of the total weight $w_{xh} + w_{xu} + w_{xp} + w_{xd}$ contributed by each x in P_d is the penalty for predicting x to be dangerous. The resulting weights $w_{x\alpha}$ for each outcome α appear on the right side of Table 1. Suppose, for instance, that the predictions are given by f_1. The penalty incurred for each x is shown in boldface in the middle of the table. The penalty for each x is the same as the negative of the sum of the boldface weights that appear at the end of that row. It is clear that finding a maximal-weight

V-partition over these weights is equivalent to finding predictions that minimize the penalty.

THEOREM 2. *Given an instance $\mathcal{M} = (X, V, c, \mu)$ of the monotone approximation problem and an arbitrary V-monotone function $f : X \longmapsto V$, let $\mathcal{P}_f = \{P_\alpha | \alpha \in V\}$ be the V-partition of the graph $G_\mathcal{M}$, defined by* (3.2)*. Then*

$$\varepsilon\,[f] = -w\,(\mathcal{P}_f)\,, \tag{3.7}$$

where $w_{x,\alpha}, \alpha \in V$ is the unique solution of (3.6) *for every $x \in X$ and where $w(\mathcal{P}_f)$ is defined by* (3.5).

Proof. We prove (3.7) as follows.

$$\begin{aligned} w\,(\mathcal{P}_f) &= \sum_{\alpha\in V}\sum_{\beta\succeq\alpha}\sum_{x\in P_\beta} w_{x,\alpha} \\ &= \sum_{\alpha\in V}\sum_{\beta\succeq\alpha}\sum_{f(x)=\beta} w_{x,\alpha} \\ &= \sum_{x\in X}\sum_{\alpha\preceq f(x)} w_{x,\alpha} \\ &= \sum_{x\in X}\left(-\sum_{\beta\in V} c_{f(x),\beta}\,\mu\,(x,\beta)\right) \\ &= -\sum_{x\in X}\sum_{\beta\in V} c_{f(x),\beta}\,\mu\,(x,\beta) \\ &= -\varepsilon\,[f]\,. \end{aligned} \tag{3.8}$$

Here, the second line comes from the first one by the definition of $\mathcal{P}_f$, the third line comes by a simple rearrangement, and the fourth equation is obtained by applying (3.6). For the last equation, we simply applied (2.1). □

COROLLARY 1. *For a given instance $\mathcal{M} = (X, V, c, \mu)$ of monotone approximation, let $w = (w_{x,\alpha} | x \in X, \alpha \in V)$ be determined by* (3.6)*. Then finding the best V-monotone fit to $\mathcal{M}$ is equivalent to finding the maximum-weight V-partition in the associated graph $G_\mathcal{M}$.*

4. Aligned partial orders.

In this section, let us consider aligned partial orders $(V, \prec)$, i.e., partial orders satisfying conditions (C1) and (C2).

Let us first prove that this property of a partial order can be recognized in polynomial time.

LEMMA 1. *Given a partial order $(V, \succ)$, one can either find a labeling $V = \{\nu_0, \nu_1, \ldots, \nu_v\}$ satisfying conditions* (C1) *and* (C2) *or conclude that $(V, \succ)$ is not aligned in $O(|V|^2)$ time.*

Proof. For an $\alpha \in V$, let $S_\alpha = \{\beta | \beta \prec \alpha\}$. Furthermore, let $\gamma \in V$ such that $|S_\gamma|$ is maximal.

We show that $(V, \succ)$ is aligned if and only if the induced relation $(V \setminus S_\gamma, \succ)$ is an empty relation and the induced partial order $(S_\gamma, \succ)$ is aligned. The recognition algorithm and its complexity follows immediately from this.

Let us first assume that $(V \setminus S_\gamma, \succ)$ is an empty relation and that the induced partial order $(S_\gamma, \succ)$ is aligned. Let $S_\gamma = \{\nu_0, \ldots, \nu_{|S_\gamma|-1}\}$ be an appropriate labeling of the elements of S_γ satisfying conditions (C1) and (C2). It is easy to verify that, by defining $\nu_{|S_\gamma|} = \gamma$ and taking an arbitrary labeling $\{\nu_{|S_\gamma|+1}, \dots, \nu_v\} = V \setminus (S_\gamma \cup \{\gamma\})$, we can obtain a labeling of V that satisfies conditions (C1) and (C2).

To see the converse direction, let us assume that $(V, \succ)$ is an aligned partial order and the labeling $V = \{\nu_0, \dots, \nu_v\}$ satisfies conditions (C1) and (C2). Let k be the largest index for which $\nu_{k-1} \prec \nu_k$. Then $S_{\nu_k} = \{\nu_j | j = 0, \dots, k-1\}$ follows from condition (C2). Furthermore, condition (C1) implies that there are no indices $j > i \geq k$ such that $\nu_j \succ \nu_i$. It follows then that $S_{\nu_i} \subseteq S_{\nu_k}$ for any $i \neq k$, and hence $|S_{\nu_k}|$ is maximal. It is also clear that the induced partial order $(S_{\nu_k}, \succ)$ is also aligned (with the same labeling) and that the induced partial order $(\{\nu_k, \dots, \nu_v\}, \succ)$ is an empty relation.

The only thing left to show is that the claim remains true for any other index k' for which $|S_{\nu_{k'}}|$ is also maximal. Let us observe that, in this case, $k' > k$ and $S_{\nu_{k'}} = S_{\nu_k}$ follow from properties (C1) and (C2). Therefore, by interchanging the elements ν_k and $\nu_{k'}$, we obtain another good labeling of V, which proves the lemma. □

We show next that interval orders are aligned.

LEMMA 2. *Interval orders are aligned.*

Proof. Let us recall (see [9]) that $(V, \prec)$ is an interval order, by definition, if there are real intervals $[a_\alpha, b_\alpha] \subseteq \mathbb{R}$ for $\alpha \in V$ such that $\alpha \prec \beta$ if and only if $b_\alpha < a_\beta$. Let us consider such a set of intervals realizing $(V, \succ)$.

It is clear that

$$S_\alpha = \{\beta | b_\beta < a_\alpha\}.$$

It follows then that $S_\alpha \subseteq S_\beta$ if $a_\alpha \leq a_\beta$; therefore $|S_\alpha|$ is maximal if and only if a_α is maximal. Let $\alpha \in V$ be such that a_α is maximal. Then, for every $\beta \in V$, either $b_\beta < a_\alpha$ (in which case, $\beta \in S_\alpha$) or $a_\beta \leq a_\alpha \leq b_\beta$. This implies that there is a common point a_α of all intervals associated to the elements of V that are not in S_α, and hence there cannot be any relation between these elements.

Since $(S_\alpha, \succ)$ is again an interval order and since $|S_\alpha| < |V|$, we can continue, and the recognition algorithm described in the previous lemma will not fail. □

It is easy to see that aligned orders are not necessarily interval orders. For example, let us consider $V = \{\nu_0, \nu_1, \nu_2, \nu_3, \nu_4\}$, in which $\nu_0 \prec \nu_2, \nu_0 \prec \nu_4, \nu_1 \prec \nu_2$, and $\nu_1 \prec \nu_3$. This partial order is clearly aligned, but it is not interval, since neither $\nu_0 \prec \nu_3$ nor $\nu_1 \prec \nu_4$.

5. Maximal V-partitions via maximal closures. In this section, we show that if $(V, \prec)$ is an aligned partial order, then a maximal weight V-partition can be determined in polynomial time.

Let $G = (N, A)$ be a given directed graph, let $(V, \prec)$ be an aligned order, and let $V = \{\nu_0, \dots, \nu_v\}$ be a labeling of the elements of V satisfying conditions (C1) and (C2). Furthermore, let $w_{x,\alpha}$ be given reals for $x \in N$ and $\alpha \in V$.

It follows then that, for any V-partition, $\mathcal{P} = \{P_\alpha | \alpha \in V\}$ and, for any index i for which $\nu_{i-1} \prec \nu_i$,

$$\hat{P}_{\nu_i} = \bigcup_{j=i}^{v} P_{\nu_j}. \tag{5.1}$$

In the example, we can let $\{\nu_0, \nu_1, \nu_2, \nu_3\} = \{h, u, p, d\}$ as an appropriate labeling. As we can observe easily, this partial order is not only aligned but is also an interval order.

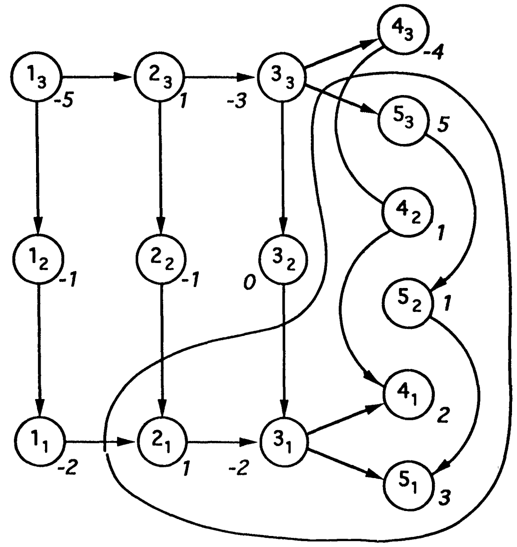


FIG. 2. *Maximal closure formulation of an approximation problem.*

Let $k_{\mathcal{P}}(x)$ be defined for any $x \in N$ as the index of $f_{\mathcal{P}}(x)$ in this linear order, where $f_{\mathcal{P}}$ is again the function defined by (3.4). In other words, $f_{\mathcal{P}}(x) = \nu_{k_{\mathcal{P}}(x)}$ for every $x \in N$.

Let us define a directed graph, $\hat{G} = (\hat{N}, \hat{A})$ associated with G and $(V, \prec)$, as follows. The graph $\hat{G}$ contains $v|N|$ nodes, and each node $x \in N$ has v "relatives" in $\hat{N}$, denoted by $x_1, x_2, \ldots, x_v$. We include the arcs (x_i, y_i) for $(x, y) \in A$ and for indices i with $\nu_{i-1} \prec \nu_i$ in the partially ordered set $(V, \prec)$. We also add the arcs (x_i, x_{i-1}) to $\hat{A}$ for every $x \in N$ and $i = 2, \ldots, v$. In other words, we place v copies of the node set N on top of each other and we include all the arcs of G in the ith level, if ν_{i-1} and ν_i are comparable; otherwise the ith level is an empty graph. In addition to these, we include a directed arc from each vertex to its lower relative, too. The graph $\hat{G}$ associated with the graph of Fig. 1 appears in Fig. 2.

We define real weights $\hat{w}_{x_i}$ associated with every vertex $x_i \in \hat{N}$ as follows:

$$\hat{w}_{x_i} = \left(\sum_{\beta \preceq \nu_i} w_{x,\beta} - \sum_{\beta \preceq \nu_{i-1}} w_{x,\beta} \right) \quad \text{for } i = 1, \ldots, v, \text{ and } x \in N. \tag{5.2}$$

The weights $\hat{w}_{x_i}$ for our example also appear in Fig. 2.

THEOREM 3. *There is a one-to-one correspondence between the V-partitions of G and the terminal sets of $\hat{G}$. Furthermore, if $\mathcal{P} = \{P_\alpha | \alpha \in V\}$ is a V-partition of G, and if $\hat{C}$ is the corresponding terminal set of $\hat{G}$, then*

$$w(\mathcal{P}) = \sum_{x \in N} w_{x,\nu_0} + \hat{w}\left(\hat{C}\right), \tag{5.3}$$

where $\hat{w}$ is the vector defined by (5.2).

Since $\sum_{x \in N} w_{x,\nu_0}$ is a constant, it follows that finding a maximal V-partition in G is equivalent with finding a maximum-weight terminal set in $\hat{G}$.

Proof. Let us consider first a V-partition of $G, \mathcal{P} = \{P_\alpha | \alpha \in V\}$, and define a subset of the nodes of $\hat{G}$ as follows:

$$\hat{C}_{\mathcal{P}} = \bigcup_{i=1}^{v} \left\{ x_i \,\middle|\, x \in \bigcup_{j=i}^{v} P_{\nu_j} \right\}. \tag{5.4}$$

We show that $\hat{C}_{\mathcal{P}}$ is a terminal set in $\hat{G}$.

For $x_i \in \hat{C}_{\mathcal{P}}$, it follows that $x_{i-1} \in \hat{C}_{\mathcal{P}}$ by definition (5.4). Let us consider an arc $(x_i, y_i) \in \hat{A}$ of $\hat{G}$, where $x_i \in \hat{C}_{\mathcal{P}}$. By the definition of $\hat{G}, (x_i, y_i) \in \hat{A}$ only if $\nu_{i-1} \prec \nu_i$. Therefore it follows from (5.1) and (3.1) that $\bigcup_{j=i}^{v} P_{\nu_j}$ is a terminal set of G. By definition, $x_i \in \hat{C}_{\mathcal{P}}$ implies that $x \in P_{\nu_j}$ for some $j \geq i$, i.e., that $x \in \bigcup_{j=i}^{v} P_{\nu_j}$. From $(x_i, y_i) \in \hat{A}$, it also follows that $(x, y) \in A$ and therefore $y \in \bigcup_{j=i}^{v} P_{\nu_j}$, thus implying that $y_i \in \hat{C}_{\mathcal{P}}$.

Conversely, let us consider a terminal set $\hat{C}$ of $\hat{G}$. Let $P_{\nu_0} = \{x | x_1 \notin \hat{C}\}$, let $P_{\nu_i} = \{x | x_i \in \hat{C}$ but $x_{i+1} \notin \hat{C}\}$, for $i = 1, 2, \ldots, v-1$, and let $P_{\nu_v} = \{x | x_v \in \hat{C}\}$. We show that $\mathcal{P} = \{P_{\nu_i} | i = 0, 1, \ldots, v\}$ is a V-partition of G such that $\hat{C} \equiv \hat{C}_{\mathcal{P}}$.

Since $\hat{C}$ is a terminal set of $\hat{G}$, it follows that, for every $x \in N$, either $x_1 \notin \hat{C}$, or $x_v \in \hat{C}$, or there is a unique $0 \leq i < v$ such that $x_i \in \hat{C}$ but $x_{i+1} \notin \hat{C}$. Therefore $\mathcal{P}$ is a partition of N.

Let us now consider an arbitrary arc $(x, y) \in A$ of G for which $x \in P_{\nu_i}$ and $y \in P_{\nu_j}$, and let us assume that $\nu_j \prec \nu_i$. Then, on the one hand, $x_i \in \hat{C}$, but $x_{j+1} \notin \hat{C}$, by the above definition of $\mathcal{P}$. On the other hand, by property (C1), there must be an index k such that $j \leq k, k+1 \leq i$, and $\nu_k \prec \nu_{k+1}$. Then the arc (x_{k+1}, y_{k+1}) belongs to the graph $\hat{G}$, and thus the vertices $x_i, x_{i-1}, \ldots, x_{k+1}, y_{k+1}, y_k, \ldots, y_{j+1}$ form a directed path from x_i to x_{j+1} in $\hat{G}$, which starts in the terminal set $\hat{C}$ and ends outside of it. This contradiction shows that $\nu_j \not\prec \nu_i$, i.e., that $k_{\mathcal{P}}(y) \not\prec k_{\mathcal{P}}(x)$ for any arc $(x, y) \in A$ of G, and thus it proves that $\mathcal{P}$ is a V-partition of G.

Finally, (5.3) follows easily by (5.2) and (5.4):

$$\begin{aligned} \hat{w}\left(\hat{C}\right) &= \sum_{x_i \in \hat{N}} \hat{w}_{x_i} \\ &= \sum_{x \in N} \sum_{i=1}^{k_P(x)} \hat{w}_{x_i} \\ &= \sum_{x \in N} \left(\sum_{\alpha \preceq f_{\mathcal{P}}(x)} w_{x,\alpha} - w_{x,\nu_0} \right) \\ &= w(\mathcal{P}) - \sum_{x \in N} w_{x,\nu_0}. \qquad \square \end{aligned} \tag{5.5}$$

The maximal closure for Fig. 2 is encircled in the figure. Note that this solution corresponds to the function f_1 defined in Table 1. Thus f_1 is a best approximation to the data.

We conclude by indicating in Table 2 the possible values of $f_1(x)$ for both observed and unobserved values of x. The observed values are in boldface. Note, for instance, that $f_1(N, U)$ cannot be d, since $(N, U) \preceq (I, U)$, $f_1(I, U) = u$, and $d \succ u$. In this example, the value of $f_1(x)$ is completely determined for only one unobserved x, namely, $x = (P, P)$.

TABLE 2
Function values consistent with an optimal approximation.

x	$f_1(x)$
$(\mathbf{N},\mathbf{N})$	$\mathbf{h}$
(U,N)	h,u,p
(N,U)	h,u,p
$(\mathbf{I},\mathbf{N})$	$\mathbf{u}$
(N,I)	h,u,p,d
(U,U)	h,u,p
$(\mathbf{I},\mathbf{U})$	$\mathbf{u}$
(P,N)	u,p,d
(U,I)	h,u,p,d
(I,I)	u,p,d
(N,P)	h,u,p,d
$(\mathbf{P},\mathbf{U})$	$\mathbf{p}$
(P,I)	u,p,d
(U,P)	h,u,p,d
$(\mathbf{I},\mathbf{P})$	$\mathbf{d}$
(P,P)	d

Acknowledgments. We thank the referees for the many helpful comments and suggestions made on an earlier version of this paper. We thank Dr. Paul B. Kantor of the School of Communication, Information and Library Studies, Rutgers University for kindly providing us with the example on carcinogenecity of chemicals, and Dr. Alex Kogan of Rutgers Center for Operations Research, Rutgers University for the helpful remarks and questions on partial ordered sets.

REFERENCES

[1] M. L. BALINSKI, *On a selection problem,* Management Sci., 17 (1970), pp. 230–231.

[2] R. E. BARLOW, D. J. BARTHOLOMEW, D. J. BREMNER, AND H. D. BRUNK, *Statistical Inference under Order Restrictions,* John Wiley, New York, 1972.

[3] M. J. BEST AND N. CHAKRAVARTI, *Active set algorithms for isotonic regression; A unifying framework,* Math. Programming, 47 (1990), pp. 425–439.

[4] E. BOROS, P. L. HAMMER, AND J. N. HOOKER, *Boolean regression,* Research Report RRR 19-91, RUTCOR, Rutgers University, New Brunswick, NJ, 1991.

[5] E. BOROS AND R. SHAMIR, *Convex Separable Minimization over Partial Order Constraints,* Research Report RRR 27-91, RUTCOR, Rutgers University, New Brunswick, NJ, 1991.

[6] C. CARTER AND J. CATLETT, *Assessing credit card applications using machine learning,* IEEE Expert, Fall 1987, pp. 71–79.

[7] Y. CRAMA, P. L. HAMMER, AND T. IBARAKI, *Cause-effect relationships and partially defined Boolean functions,* Ann. Oper. Res., 16 (1988), pp. 299–325.

[8] B. FAALAND, K. KIM, AND T. SCHMITT, *A new algorithm for computing the maximal closure of a graph,* Management Sci., 36 (1990), pp. 315–331.

[9] P. C. FISHBURN, *Interval Orders and Interval Graphs,* John Wiley, New York, 1985.

[10] P. L. HAMMER AND S. RUDEANU, *Boolean Methods in Operations Research and Related Areas,* Springer-Verlag, Berlin, Heidelberg, New York, 1968.

[11] P. HANSEN, *Quelques Approches de la Programmation Non Lineaire en variables* 0-1, Conference on Mathematical Programming, Bruxelles, 1974.

[12] R. F. HAUCK AND J. M. MALONE, *Optimal Open Pit Mine Contours with Optimal Depletion Scheduling: Theory and Methodology,* U.S. Steel Corp., 1969.

[13] J. W. MAMER AND S. A. SMITH, *Optimizing field repair kits based on job completion rate,* Management Sci., 28 (1982), pp. 1328–1333.

[14] W. L. MAXWELL AND J. A. MUCKSTADT, *Establishing consistent and realistic reorder intervals in production-distribution systems,* Oper. Res., 33 (1985), pp. 1316–1341.

[15] J. D. MURCHLAND, *Rhys's combinatorial station selection problem,* London Graduate School of Business Studies, Transport Network Theory Unit, Report LBS-TNT-68, London, 1968.

[16] J.-C. PICARD, *Maximal closure of a graph and applications to combinatorial problems,* Management Sci., 22 (1976), pp. 1268–1272.

[17] J.-C. PICARD AND M. QUEYRANNE, *Selected applications of minimum cuts in networks,* INFOR, 20 (1982), pp. 394–422.

[18] ———, *Integer Minimization of a Separable Convex Function Subject to Variable Upper Bound Constraints,* Research Report, University of British Columbia, Vancouver, BC, April 1985.

[19] R. ROUNDY, *A 98% effective lot-sizing rule for a multi-product, multi-stage production/ inventory system,* Math. Oper. Res., 11 (1986), pp. 699–727.

[20] J. M. W. RHYS, *A selection problem of shared fixed costs and network flows,* Management Sci., 17 (1970), pp. 200–207.

[21] M. J. SHAW AND J. A. GENTRY, *Inductive learning for risk classification,* IEEE Expert, February 1990, pp. 47–53.

[22] K. A. SPACKMAN, *Learning categorical decision criteria in biomedical domains,* in Proc. of the Fifth International Conference on Machine Learning, 1988, pp. 36–46.

[23] H. M. WEINGARTNER, *Capital budgeting of interrelated projects: Survey and synthesis,* Management Sci., 12 (1966), pp. 485–516.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 544–553, November 1994

A CODING APPROACH TO SIGNED GRAPHS*

PATRICK SOLÉ† AND THOMAS ZASLAVSKY‡

Abstract. The cocycle code of an undirected graph Γ is the linear span over $\mathbf{F}_2$ of the characteristic vectors of cutsets. (If Γ is complete bipartite, this is the generalized Gale–Berlekamp code.) The natural bijection between the cosets of this code and the switching classes of signed graphs based on Γ is used to show that the number of such classes is equal to the number of even-degree subgraphs of Γ in both the labeled and unlabeled cases and to improve by coding theory previous bounds on $D(\Gamma)$, the maximum line index of imbalance of signings of Γ. Bounds on $D(\Gamma)$ are obtained in terms of the genus of Γ and on the number of unlabeled even-degree subgraphs in terms of $D(\Gamma)$. Numerous examples are treated, including the "grid" (or "lattice") graphs that are of interest in the Ising model of spin glasses.

Key words. signed graph, line index of imbalance, frustration, cutset code, cocycle code, covering radius, genus, spin glass, Ising model, Gale–Berlekamp code, switching class, even-degree subgraph

AMS subject classifications. primary 05C35; secondary 94B25, 82B20

1. Introduction.

Graphs with signed edges, called *signed graphs,* were introduced in the 1950s [12] in connection with a problem in attitudinal psychology [7] and have since often been rediscovered, notably in the theory of spin glasses [31]. The fundamental feature of a signed graph is the list of circuits that are *positive,* that is, have an even number of negative edges. We call two signed graphs *switching equivalent* if they have the same base graph and the same positive circuits. The equivalence classes (called *switching classes*) of signings of a graph correspond to the additive cosets of the binary cocycle space of Γ; by considering this space as a linear code (the *cocycle code* or *cutset code*), we can use coding theory to obtain results on two problems about signed graphs.

A signed graph is called *balanced* if every cycle is positive. One measure of the degree to which a signed graph fails to be balanced is the smallest number of edges whose negation produces a balanced signed graph. This quantity is known as the *line index of imbalance* or *frustration index.* (The former name is from [13]; the concept, in different formulations, originates in [1] and [13] and later from spin glass theory [3, §2.2]—whence comes the term "frustration"—and possibly elsewhere.) The question is how large the index can be, given Γ. We obtain new upper and lower bounds on the maximum index, improving the results of [2], by observing that it equals the covering radius of the cutset code.

Using an approach to imbalance in signed planar graphs first developed in [15] and our recent results on the covering radius of cycle codes [29], we obtain lower bounds on the line index of imbalance of a signed graph based on a graph of known genus.

* Received by the editors December 4, 1989; accepted for publication (in revised form) September 1, 1993.

† The School of Computer and Information Science, Syracuse University, Syracuse, New York 13244-4100. Present address, Centre National de la Recherche Scientifique, I3S, bât. 4, 250 rue Albert Einstein, Sophia Antipolis, 06560 Valbonne, France.

‡ Department of Mathematical Sciences, State University of New York at Binghamton, Binghamton, New York 13902-6000. The research of this author was supported by National Science Foundation grant DMS-8808239.

The number of switching classes of signings of a graph Γ is known to equal the number of even-degree subgraphs of Γ. This is trivial if Γ is labeled, a theorem of Wells if Γ is unlabeled [32]. We observe that Wells's theorem is precisely the application to cutset codes of a standard automorphism property of a linear code. Then we use the cutset code to obtain a lower bound on the number of unlabeled even-degree subgraphs.

2. The cutset code of a graph. We let Γ denote a finite, loopless graph with vertex set $V = V(\Gamma)$ and edge set $E = E(\Gamma) = \{e_1, e_2, \dots, e_m\}$. We write $n = |V|, m = |E|$, and $c =$ the number of connected components of Γ. A *cocycle* or *cutset* is an edge set that consists of all edges having one endpoint in some set $X \subseteq V$ and one endpoint not in X. Under the operation of set sum (i.e., symmetric difference), the cutsets form a subspace of the binary vector space of all subsets of E. Replacing subsets S of E by their characteristic vectors $x_S \in \mathbf{F}_2^m$ gives a subspace of $\mathbf{F}_2^m$, which we call the *cutset* or *cocycle code,* $C^*(\Gamma)$. The dual code is the *cycle code* $C(\Gamma)$, defined as the linear span of the characteristic vectors of circuits of Γ. As is well known, both codes have length m; their dimensions are $n - c$ for the cocycle code and $m - n + c$ for the cycle code.

The *cosets* of a linear code C are the additive cosets of C in $\mathbf{F}_2^m$. The *weight* $w(x)$ of an element x of $\mathbf{F}_2^m$ is the number of 1s it contains. The *distance* from x to y in $\mathbf{F}_2^m$ is $d(x,y) = w(x-y)$. The *minimum weight* of a coset $x + C$ is $\min_{y \in C} w(x+y)$. The *covering radius* $R(C)$ is the largest minimum weight of a coset taken over all cosets.

3. Signed graphs and switching classes. A *signed graph* is a pair (Γ, σ), where σ is a function from the edges of Γ into $\{-1, +1\}$. A signed graph (Γ, σ') is said to be obtained by *switching* (Γ, σ) at $X \subseteq V$ if

$$\sigma'(xy) = \begin{cases} \sigma(xy) & \text{if } x, y \in X \text{ or } x, y \in V \backslash X, \\ -\sigma(xy) & \text{if } x \in X \text{ and } y \notin X \text{ or } x \notin X \text{ and } y \in X, \end{cases}$$

where xy denotes an edge whose endpoints are x and y. Two signed graphs are said to be *switching equivalent* if they are based on the same graph and are exchanged by switching at some X. It is easy to see that switching equivalence is an equivalence relation. It is also easy to show that two signings of Γ are switching equivalent if and only if they have the same positive circuits (see [33, Prop. 3.2], for instance).

Switching is derived ultimately from [16], where it was employed to study two-graphs, which are, in effect, signed graphs (K_n, σ) on the complete graph K_n. The precise definition is as follows: a *two-graph* on n vertices is a class of three-element subsets of an n-set V such that each four-element subset of V contains an even number of members of the class. The original article on two-graphs is [30]. An excellent exposition of this line of development is [24].

A simple but crucial lemma is the following.

LEMMA 1. *The mapping* $+1 \to 0, \quad -1 \to 1$ *induces one-to-one correspondences* (i) *between signed graphs based on* Γ *and elements of* $\mathbf{F}_2^m$, *and* (ii) *between switching classes of signed graphs based on* Γ *and cosets of* $C^*(\Gamma)$.

Proof. (i) is obvious. (ii) Switching at X in $(+1, -1)$ notation is equivalent to adding in (0, 1) notation the cocycle determined by X and its complement. Thus a switching class is, in $(0,1)$ notation, a coset of $C^*(\Gamma)$. □

4. Covering radius and index of imbalance. It is clear that a signed graph is balanced if and only if it is switching equivalent to the all-plus signed graph $(\Gamma, +1)$. The smallest number $d(\Gamma, \sigma)$ such that (Γ, σ) can be changed into a balanced graph by changing the sign of $d(\Gamma, \sigma)$ edges is called the *line index of imbalance* of (Γ, σ). The largest possible line index of imbalance over all signings of a graph Γ, call it $D(\Gamma)$, was investigated in [2]. These authors proved inter alia that

$$\frac{m}{2} - \sqrt{nm} \leq D(\Gamma) \leq \frac{m}{2} \tag{1}$$

and

$$D(K_{t,t}) \leq \frac{t^2}{2} - c_0 t^{3/2} \tag{2}$$

for some constant c_0 that can be taken as $c_0 = \pi/480$.

We can improve these results by using known facts on the covering radius of codes (some of which were originally expressed in terms of ±1 matrices). Theorem 1 improves the lower bound in (1). Our proof is, moreover, conceptually simpler than the proof of (1). Formula (6) sharpens (2), and Theorem 2 shows that the quantity $m/2 - D(\Gamma)$ can be bounded away from zero more generally than is done in (2).

THEOREM 1. *It is the case that*

$$\frac{m}{2} - \sqrt{\frac{\ln 2}{2}} \sqrt{m(n-c)} \leq D(\Gamma). \tag{3}$$

We mention that (3) is derived in [26] from a slightly stronger formula, which in our notation is

$$D(\Gamma) \geq mH^{-1}\left(1 - \frac{n-c}{m}\right), \tag{4}$$

where H is the binary entropy function (H_2 in [17, pp. 308–309]).

THEOREM 2. *If* Γ *is simple and bipartite,*

$$D(\Gamma) \leq \frac{m - \sqrt{m}}{2}. \tag{5}$$

THEOREM 3. [5, p. 266, Thm.]. *We have*

$$\frac{t^2}{2} - \frac{t^{3/2}}{2} + o(t^{3/2}) \leq D(K_{t,t}) \leq \frac{t^2}{2} - \frac{t^{3/2}}{\sqrt{2\pi}} + o(t^{3/2}). \tag{6}$$

The error term in the upper bound of Theorem 3 was reduced to $O(t^{1/2})$ in [9, Cor. 3 and Rem.].

For the proofs, we first need a lemma.

LEMMA 2.

(i) *The line index of imbalance of a signed graph based on* Γ *is equal to its minimum weight as a coset of* Γ.

(ii) $D(\Gamma)$ *is equal to the covering radius of* $C^*(\Gamma)$.

Proof. (i) is clear by translating $(+1, -1)$ notation into $(0, 1)$ notation. (ii) is clear from (i), since the covering radius of a code is the largest possible minimum weight of a coset. □

Proof of Theorem 1. For any $[N, K]$ code, a simple consequence of the sphere-covering bound is $R \geq N/2 - \sqrt{\frac{1}{2}NK \ln 2}$ (stated in [20]; the proof is given in [26, Thm. 2]). Here $N = m, K = n - c$, and $R = D(\Gamma)$. □

Proof of Theorem 2. Furthermore, if a binary code is of strength 2 and contains the all-one vector, then $R \leq (N - \sqrt{N})/2$ [14, Thm. 3]. Here the strength is 2 because the minimum distance of the cycle code is ≥ 3. $C^*(\Gamma)$ contains the all-one vector if and only if Γ is bipartite. □

Note that, for any graph, the bound $D(\Gamma) \leq m/2$ of [2] is also implied by the results of [14].

Proof of Theorem 3. We indicate how the quantity evaluated in [5], which is the minimum number of -1s in line negations of a ± 1-matrix (of order $p \times q$, say), maximized over all such matrices, equals $D(K_{p,q})$. Let $V(K_{p,q}) = X \cup Y$, where $X = \{x_1, \ldots, x_p\}$ and $Y = \{y_1, \ldots, y_q\}$. An entry a_{ij} in the matrix is the sign of the edge $x_i y_j$. Switching a vertex, say x_i, corresponds to negating a line, in this case, row i. □

Interest in $D(K_{p,q})$, although not by that name, was aroused by the Gale–Berlekamp switching game [5], [8]–[10] and the corresponding binary code, which happens to be $C^*(K_{10,10})$. The code $C^*(K_{p,q})$ is usually described as the p by q array code whose codewords are sums of matrices of the form

$$\begin{bmatrix} & 1 & \\ 0 & \vdots & 0 \\ & 1 & \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} & 0 & \\ 1 & \cdots & 1 \\ & 0 & \end{bmatrix}.$$

This is a $[pq, p+q-1]$ code, which might be called a *generalized Gale–Berlekamp code* because it corresponds to the p by q Gale–Berlekamp switching game. Identifying rows (respectively, columns) with the first set (respectively, second set) of the bipartition of $K_{p,q}$, we see that the code is $C^*(K_{p,q})$. Indeed, switching at a point in $K_{10,10}$ is the same as switching a commutator in the Gale–Berlekamp switching game.

Theorems 1 and 3 suggest that the best general asymptotic lower bound on $D(\Gamma)$ has the form

$$\frac{m}{2} - c_1\sqrt{mn} + o\left(\sqrt{mn}\right) \leq D(\Gamma),$$

where c_1 is a positive constant lying between $\sqrt{\frac{1}{2}\ln 2} \approx .59$ and $1/2\sqrt{\pi} \approx .28$. We would like to know the exact value of c_1, both for all graphs and for narrower classes, especially for classes in which m/n^2 is bounded away from zero and for the class of k-connected graphs, where $k \geq 1$.

For the latter, the graphs $K_{p,q}$ with $k \leq p \leq q$ are relevant. It is known [9, Thm. 3], [10, p. 396, Ex. b] that

$$(7) \qquad D(K_{p,q}) \leq \frac{pq}{2} - \frac{pq}{2^p}\binom{p-1}{\left\lfloor\frac{p-1}{2}\right\rfloor} \approx \frac{1}{\sqrt{2\pi}}\sqrt{mn} \quad \text{if } q >> p.$$

Exact values are known when $p \leq 5$. It is easy to prove that

$$D(K_{1,q}) = 0, \quad D(K_{2,q}) = \left\lfloor\frac{q}{2}\right\rfloor, \quad D(K_{3,q}) = \left\lfloor\frac{3q}{4}\right\rfloor$$

for all q. It is true, but difficult to prove, that

$$D(K_{4,q}) = \left\lfloor\frac{5}{4}q\right\rfloor - \alpha_4,$$

where $\alpha_4 = 0$, except $\alpha_4 = 1$ if $q \equiv 1, 4, 5 \pmod 8$, and that

$$D(K_{5,q}) = \left\lfloor \frac{25}{16} q \right\rfloor - \alpha_5,$$

where $\alpha_5 = 0$, except $\alpha_5 = 1$ if $q \equiv 2, 4, 9, 13, 15 \pmod{16}$. (The formulas for $p \leq 4$ were published in [5] without proof. That for $p = 5$ is new.) We conclude that, as $q \to \infty$ while p is fixed,

$$D(K_{p,q}) = \frac{pq}{2} - b_p\sqrt{mn} + O(1),$$

where $b_1 = \frac{1}{2}, b_2 = \sqrt{2}/4 \approx .35, b_3 = \sqrt{3}/4 \approx .43, b_4 = \frac{3}{8} = .375$, and $b_5 = 3\sqrt{5}/16 \approx .42$. (The values approach $1/\sqrt{2\pi} \approx .40$, by (7).)

The exact values $D(K_{t,t})$ are also known, for $t \leq 10$. They are published in [8, Table 1] as the values R_t of the covering radius of the Gale–Berlekamp codes $C^*(K_{t,t})$.

Concerning Theorem 2, we conjecture that a somewhat similar upper bound holds for all simple graphs.

CONJECTURE 1. $D(\Gamma) \leq m/2 - (n-c)/4$.

The conjectured bound is sharp since it is essentially equal to D for K_n, since $D(K_n) = \lceil (n-1)/2)^2 \rceil$. (This was stated implicitly in [1, Thm. 14]. It was proved in [21].) Conjecture 1 is stronger than Theorem 2 for most bipartite graphs, including all those in which $q \geq p + 2\sqrt{p} + 1$.

For graphs with large girth, stronger upper bounds than those of (1) and Theorem 2 follow from the generalized Norse bounds of [28]. We let $\mu_k(m)$ denote the kth centered moment of a $B(m, \frac{1}{2})$ binomial probability distribution. Some small values, taken from [28], are

$$\mu_2(m) = \frac{m}{2^2}, \quad \mu_4(m) = \frac{3m^2 - 2m}{2^4}, \quad \mu_6(m) = \frac{15m^3 - 30m^2 + 16m}{2^6}.$$

THEOREM 4. *If Γ has girth $g \geq 2s + 2$, where s is a positive integer, then*

$$D(\Gamma) \leq \frac{m}{2} - \left(2^{1-1/(2s+1)} - 1\right) \mu_{2s}(m)^{1/2s}.$$

If Γ is bipartite, then

$$D(\Gamma) \leq \frac{m}{2} - \mu_{2s}(m)^{1/2s}.$$

Proof. The proof follows from [28, Thms. 6 and 1]. □

When $s = 1$, the latter part of this theorem is Theorem 2.

The next result is an asymptotic version of Theorem 4.

THEOREM 5. *If Γ has girth $g \geq 2s + 2$, where s is a fixed positive integer, then, as $m \to \infty$, we have*

$$D(\Gamma) \leq \frac{m}{2} - \frac{\sqrt{m}}{2}\left(2^{1-1/(2s+1)} - 1\right)\left(\frac{(2s)!}{s!}\right)^{1/2s} + o\left(\sqrt{m}\right)$$

and, if Γ is bipartite,

$$D(\Gamma) \leq \frac{m}{2} - \frac{\sqrt{m}}{2}\left(\frac{(2s)!}{s!}\right)^{1/2s} + o\left(\sqrt{m}\right).$$

Proof. The proof follows from [28, Thms. 7 and 3]. □

For general graphs with fixed minimum girth, better estimates are attainable. For instance, from [28, Thms. 4 and 5], we obtain Theorem 6.

THEOREM 6. *If* Γ *has girth* ≥ 4, *then*

$$D(\Gamma) \leq \frac{m}{2} - f_1 \frac{\sqrt{m}}{2},$$

where $f_1 \approx .68$.

If Γ *has girth* $g \geq 6$, *then*

$$D(\Gamma) \leq \frac{m}{2} - h_2 \mu_4 (m)^{1/4},$$

where $h_2 \approx .82$.

The exact values of f_1 and h_2 can be obtained from [28].

The bounds of Theorem 6 are stronger than those of Theorem 4 evaluated at $s = 1$ and 2. However, they are still probably rather weak. (Since they are so general, applying to all binary codes, that is not surprising. The same remark applies to all general results of this section, although they are the best known bounds on $D(\Gamma)$.) For instance, the Petersen graph P has girth 5. Theorem 6 gives $D(P) \leq 5.7$, but it can be shown that $D(P) = 3$. For an example falling under Theorem 2 (i.e., the case where $s = 1$ and Γ is bipartite of Theorem 4), see Example 10.

5. Embedded graphs. In this section, all graphs are connected. For many planar graphs $\Gamma, D(\Gamma)$ is easy to compute because of duality. Let Γ^* be the planar dual of Γ. Then $C^*(\Gamma) = C(\Gamma^*)$; hence $D(\Gamma) = R(C(\Gamma^*))$. However, this equals $\lfloor \frac{1}{2} n^* \rfloor$ (where $n^* = |V(\Gamma^*)|$ = the number of faces of a planar embedding of Γ, that is, $m + 2 - n$) for many graphs, for instance, if Γ^* is Hamiltonian or if it is connected, k-regular, and k-edge-connected [29].

As far as we know, the first to apply planar duality to find the imbalance of signed graphs were Katai and Iwai [15, §4]. They used minimum T-joins in Γ^*, for suitable T, to calculate $d(\Gamma, \sigma)$. (T-joins and their connection to the cycle code are explained in [29].) Previously, $d(\Gamma, -1)$ had been obtained similarly in [19], where it was regarded as the largest cut size in the plane graph Γ.

Example 1. $K_{2,q}$ is planar. Its dual graph is C_q with doubled edges, which is Hamiltonian. We recover the result of §4, $D(K_{2,q}) = \lfloor q/2 \rfloor$.

Example 2. A *wheel* W_n is the join of a circuit C_{n-1} and a vertex. It is planar and self-dual. Because it is Hamiltonian, $D(W_n) = \lfloor n/2 \rfloor$.

Example 3. A *biwheel* B_n is the join $C_{n-2} \vee \bar{K}_2$, where $\bar{K}_p$ is the edgeless graph of order p. Its planar dual is a *circular ladder* CL_{n-2}, which consists of two vertex-disjoint C_{n-2}'s, say with vertex sets $\{x_1, x_2, \ldots, x_{n-2}\}$ and $\{y_1, y_2, \ldots, y_{n-2}\}$, and edges $x_i y_i$ for all i. This is Hamiltonian (and cubic and three-edge-connected), so we have $D(B_n) = n - 2$.

Example 4. Dually, $D(CL_n) = \lfloor n/2 \rfloor + 1$ since B_{n+2} is Hamiltonian.

Example 5. The *ladder* L_n is CL_n with the edges $x_1 x_n$ and $y_1 y_n$ deleted. We see that $D(L_n) = \lfloor n/2 \rfloor$, since its dual contains the Hamiltonian spanning subgraph $P_{n-1} \vee K_1$, where P_{n-1} is a path of order $n - 1$. The latter (a *fan*) is self-dual; hence $D(P_{n-1} \vee K_1) = \lfloor n/2 \rfloor$, as well.

Example 6. The *rectangular grid* $G_{p,q}$ is of some interest as a finite spin-glass model [31], [3]. It has vertex set $\{1, 2, \ldots, p\} \times \{1, 2, \ldots, q\}$ and edge set $\{(i_1, j_1)(i_2, j_2) : |i_1 - i_2| + |j_1 - j_2| = 1\}$. So $n = pq$ and $m = 2pq - p - q$. Its planar dual $G^*_{p,q}$ is $G_{p-1,q-1}$ with an extra vertex adjacent to every outer vertex of $G_{p-1,q-1}$. Since $G^*_{p,q}$ is Hamiltonian and has $(p-1)(q-1)+1$ vertices,

$$D\left(G_{p,q}\right) = \left\lfloor \frac{pq-p-q+2}{2} \right\rfloor .$$

The bounds given by Theorems 1 and 2 are far from the true value. Theorem 1, for instance, approximately gives $D > pq(1-\sqrt{\ln 2}) \approx .17pq$ for $G_{p,q}$.

We can generalize this work to nonplanar graphs as follows: Let χ denote the largest Euler characteristic of any surface in which Γ embeds and χ_0 the maximum over orientable embedding surfaces. We call $\gamma = \frac{1}{2}(2-\chi_0)$ the *genus* and $\hat{\gamma} = 2-\chi$ the *demi-genus* of Γ. A simple use of the Euler formula yields the following lemma.

LEMMA 3. *Let* Γ^* *denote the dual of* Γ *as cellularly embedded in a compact surface of demi-genus d. Then*

$$\dim C\left(\Gamma^*\right) - \dim C^*\left(\Gamma\right) = d.$$

Proof. Subtract $\dim C(\Gamma^*) = m-n^*+1$ from $\dim C(\Gamma) = n-1$ and apply Euler's formula $n-m+n^* = 2-d$. □

Using simple results of coding theory yields the following result. We denote the covering radius of $C(\Gamma)$ by $\tau(\Gamma)$.

THEOREM 7. *We have the bounds*

$$\left\lfloor \frac{n\left(\Gamma^*\right)}{2} \right\rfloor \le \tau\left(\Gamma^*\right) \le D(\Gamma) \quad \textit{and} \quad \tau\left(\Gamma^*\right) \ge \left\lfloor \frac{D\left(\Gamma\right)}{1+\hat{\gamma}} \right\rfloor .$$

Proof. $C^*(\Gamma) \subseteq C(\Gamma^*)$, hence their covering radii satisfy the reversed relationship $\tau(\Gamma^*) \le D(\Gamma)$. However, $\tau(\Gamma^*) \ge \lfloor n(\Gamma^*)/2 \rfloor$ by [29]. The second lower bound on $\tau(\Gamma^*)$ follows from Lemma 3 and the theorem of Simonis [25]. □

COROLLARY 1. $D(\Gamma) \ge \lfloor 1 - \hat{\gamma}/2 + (m-n)/2 \rfloor$.

Proof. Substitute from Euler's formula into the first bound of Theorem 7. □

For fixed $\hat{\gamma}$, this bound is better than Theorem 1, as the following examples (taken from [4]) show.

Example 7. *The Heawood graph.* $n = 14, m = 21$, and the genus is 1. We may take $\hat{\gamma} = 2$. Corollary 1 yields $D \ge 3$, while Theorem 1 yields $D \ge 2$.

Example 8. *The Franklin graph.* $n = 12, m = 18$, and we may take $\hat{\gamma} = 2$. This gives a lower bound of 3. Theorem 1 only gives 1.

Example 9. The r-dimensional *hypercube graph* Q_r has $n = 2^r$ and $m = 2^{r-1}r$. By Theorems 1 and 2,

$$2^{r-2}r\left(1-\sqrt{\frac{4\ln 2}{r}}\right) < D\left(Q_r\right) \le 2^{r-2}r - \sqrt{2^{r-3}r}. \tag{8}$$

(Here $4\ln 2 = 2.772\ldots.$)

Computing $D(Q_r)$ exactly is surely hard, but we know

$$D\left(Q_1\right) = 0, \quad D\left(Q_2\right) = 1, \quad D\left(Q_3\right) = 3.$$

The values for $r \le 2$ are obvious. Q_3 is planar, and Q_3^* is the octahedral graph. The latter is Hamiltonian, so $D(Q_3) = \tau(Q_3^*) = \lfloor \frac{1}{2}n^* \rfloor = 3$. (From (8), we only obtain $1 \le D(Q_3) \le 4$.)

We also know that the minimum $\hat{\gamma}$ for Q_r is $2+2^{r-2}(r-4)$ if $r \geq 3$ ([23]; see [11, Thm. 3.5.8]). Corollary 1 then gives the lower bound

$$D\left(Q_r\right) \geq 2^{r-3} r=\frac{m}{4}. \tag{9}$$

For instance, $8 \leq D(Q_4) \leq 13$, which is a wide range of uncertainty, but less than that allowed by (8). It is interesting that (9) holds with equality for $r=2$ and 3. Formula (8) shows this cannot be true for $r \geq 12$.

Example 10. The *toroidal grid* $G'_{p,q}$ has been used as a spin glass model [31], [3]. It consists of $G_{p,q}$ (Example 6), together with "wraparound" edges $(i,1)(i,q)$ and $(1,j)(p,j)$. It has $n=pq, m=2pq$, and $\hat{\gamma}=2$. Thus, by Corollary 1, $D(G'_{p,q}) \geq \lfloor pq/2 \rfloor$. Theorem 7 gives the same result, since G'_{pq} is self-dual in the torus, and it is Hamiltonian so $\tau(G'_{p,q})=\lfloor pq/2 \rfloor$.

On the other hand, if $\Gamma_1 \subseteq \Gamma_2$, clearly $D(\Gamma_2) \leq D(\Gamma_1)+m_2-m_1$. Taking $G_{p,q} \subseteq G'_{p,q}$, our best estimate for the toroidal grid is therefore

$$\left\lfloor\frac{pq}{2}\right\rfloor \leq D\left(G'_{p,q}\right) \leq\left\lfloor\frac{pq+p+q+2}{2}\right\rfloor.$$

Exactly where the true value lies we do not know. It is at neither extreme, since we can show that

$$D\left(G'_{1,q}\right)=q+1 \quad \text{and} \quad D\left(G'_{2,q}\right)=q+2.$$

Theorems 1 and 2 give very poor estimates, similar to those for the rectangular grid in Example 6.

6. Switching classes and even-degree subgraphs.

Now we turn to our second problem. An *even-degree subgraph* of Γ is a spanning subgraph in which all degrees are even. In other words, it is a spanning subgraph whose edge set is an element of the cycle space $C(\Gamma)$, the dual code of $C^*(\Gamma)$. It is easy to see the next theorem by comparing the dimensions of $C(\Gamma)$ and $\mathbf{F}_2^m/C^*(\Gamma)$.

THEOREM 8. *For a labeled graph* Γ, *the number of switching classes of signings of* Γ *equals the number of even-degree subgraphs.*

What happens in the unlabeled case, specifically if we let the automorphism group of Γ act on the cosets of $C^*(\Gamma)$ and the words of $C(\Gamma)$? Call two subgraphs of Γ, or (switching classes of) signings, Γ-*isomorphic* if one is carried to the other by an automorphism of Γ. Then we have the following result.

THEOREM 9. (see [32]). *The number of* Γ-*isomorphic even-degree subgraphs of* Γ *is equal to the number of* Γ-*isomorphic switching classes of signed graphs based on* Γ.

This result was proved for $\Gamma=K_n$ in [18], then in a cohomological context in [6], and later generalized to arbitrary Γ, again with a cohomological proof [32, Thm. 1.3].

Proof. We apply to the cycle code the well-known fact [22, p. 211] that a group acting on the coordinate places of a linear code C has as many orbits on the codewords of C as on the cosets of its dual. □

We now use imbalance to obtain a lower bound on the number in Theorem 9.

THEOREM 10. *The number of* Γ-*isomorphic even-degree subgraphs of* Γ *is at least equal to* $D(\Gamma)+1$.

Proof. Since two Γ-isomorphic signed graphs based on Γ correspond to equivalent cosets of $C^*(\Gamma)$, they have the same line index of imbalance. Hence, there are at least $D(\Gamma)+1$ signed graphs based on Γ that are Γ-nonisomorphic and, from Theorem 9, at least that many even-degree subgraphs in Γ. □

COROLLARY 2. *There are at least* $\lfloor (n-1)^2/4 \rfloor + 1$ *two-graphs on* n *vertices.*

Proof. From [21], we know that $D(K_n) = \lfloor (n-1)^2/4 \rfloor$. □

Call a binary linear code *completely transitive* [27] if its covering radius equals the number of orbits of cosets under the action of its automorphism group (not counting the code itself).

COROLLARY 3. *If* $D(\Gamma)+1$ *equals the number of* $\Gamma-$*isomorphic even-degree subgraphs of* Γ, *then* $C^*(\Gamma)$ *is completely transitive.*

Proof. From Theorem 9 and Lemma 1, $D(\Gamma) + 1$ counts the number of Γ-isomorphism classes of cosets of $C^*(\Gamma)$. However, every automorphism of Γ induces a coordinate automorphism of $C^*(\Gamma)$. □

Example 11. Consider $C^*(K_4)$. Note that $K_4 = W_4$. Hence, from §5 or [21], we know that $D(K_4) = 2$. All even subgraphs of K_4 are isomorphic to either $\bar{K}_4, K_3$, or C_4. Thus we obtain a completely transitive [8, 3] code with covering radius 2.

Example 12. $C^*(K_{3,3})$ is completely transitive. A direct proof is given in [27]. All even subgraphs are isomorphic to either $\bar{K}_6, C_4$, or C_6. It is known from [5] (and easily proved) that $D(K_{3,3}) = 2$. So we have a completely transitive [9, 5] code with covering radius 2.

7. Conclusion and open problems. In this paper, we have shown how coding theory could be used to generalize certain enumerative results on two-graphs and to improve certain estimates on the line index of imbalance of signed graphs. The next question is to see if graph theory can be of some use in designing good new covering codes or decoding classical ones.

Acknowledgment. We thank H. F. Mattson, Jr. for helpful discussions.

REFERENCES

[1] R. P. ABELSON AND M. J. ROSENBERG, *Symbolic psycho-logic: A model of attitudinal cognition*, Behavioral Sci., 3 (1958), pp. 1–13.

[2] J. AKIYAMA, D. AVIS, V. CHVATAL, AND H. ERA, *Balancing signed graphs*, Discrete Appl. Math., 3 (1981), pp. 227–233.

[3] I. BIECHE, R. MAYNARD, R. RAMMAL, AND J. P. UHRY, *On the ground states of the frustration model of a spin glass by a matching method of graph theory*, J. Phys. A: Math. Gen., 13 (1980), pp. 2553–2576.

[4] J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, American Elsevier, New York, 1976.

[5] T. A. BROWN AND J. H. SPENCER, *Minimization of* ±1 *matrices under line shifts*, Colloq. Math., 23 (1971), pp. 165–171.

[6] P. J. CAMERON, *Cohomological aspects of two-graphs*, Math. Z., 157 (1977), pp. 101–119.

[7] D. CARTWRIGHT AND F. HARARY, *Structural balance: A generalization of Heider's theory*, Psychological Rev., 63 (1956), pp. 277–293; reprinted in Group Dynamics: Research and Theory, 2nd ed., D. Cartwright and A. Zander, eds., Harper and Row, New York, 1960, pp. 705–726; also reprinted in Social Networks: A Developing Paradigm, S. Leinhardt, ed., Academic Press, New York, 1977, pp. 9–25.

[8] P. C. FISHBURN AND N. J. A. SLOANE, *The solution to Berlekamp's switching game*, Discrete Math., 74 (1989), pp. 263–290.

[9] Y. GORDON AND H. S. WITSENHAUSEN, *On extensions of the Gale–Berlekamp switching problem and constants of* l_p*-spaces*, Israel J. Math., 11 (1972), pp. 216–229.

[10] R. L. GRAHAM AND N. J. A. SLOANE, *On the covering radius of codes*, IEEE Trans. Inform. Theory, IT-31 (1985), pp. 385–401.

[11] J. L. GROSS AND T. W. TUCKER, *Topological Graph Theory*, Wiley-Interscience, New York, 1987.

[12] F. HARARY, *On the notion of balance of a signed graph*, Michigan Math. J., 2 (1953–54), pp. 143–146 (addendum, ibid., preceding p. 1).

[13] ———, *On the measurement of structural balance*, Behavioral Sci., 4 (1959), pp. 316–323.

[14] T. HELLESETH, T. KLØVE, AND J. MYKKELVEIT, *On the covering radius of binary codes*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 627–628.

[15] O. KATAI AND S. IWAI, *Studies on the balancing, the minimal balancing, and the minimum balancing processes for social groups with planar and nonplanar graph structures*, J. Math. Psych., 18 (1978), pp. 140–176.

[16] J. H. VAN LINT AND J. J. SEIDEL, *Equilateral point sets in elliptic geometry*, Indag. Math., 28 (1966), pp. 335–348.

[17] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[18] C. L. MALLOWS AND N. J. A. SLOANE, *Two-graphs, switching classes and Euler graphs are equal in number*, SIAM J. Appl. Math., 28 (1975), pp. 876–880.

[19] G. I. ORLOVA AND YA. G. DORFMAN, *Finding the maximum cut in a graph*, Engrg. Cybernetics, 10 (1972), pp. 502–506.

[20] J. PACH AND J. SPENCER, *Explicit codes with low covering radius*, IEEE Trans. Inform. Theory, IT-34 (1988), pp. 1281–1285.

[21] M. PETERSDORF, *Einige Bemerkungen über vollständige Bigraphen*, Wiss. Z. Techn. Hochsch. Ilmenau, 12 (1966), pp. 257–260.

[22] W. W. PETERSON, *Error Correcting Codes*, MIT Press, Cambridge, MA, 1961.

[23] G. RINGEL, *Über drei kombinatorische Probleme am n-dimensionalen Wurfel und Wurfelgitter*, Abh. Math. Sem. Univ. Hamburg, 20 (1955), pp. 10–19.

[24] J. J. SEIDEL, *A survey of two-graphs*, Colloquio Internazionale sulle Teorie Combinatorie, Rome, 1973, Accad. Naz. Lincei, Rome, 1976, Vol. I, pp. 481–511.

[25] J. SIMONIS, *A generalization of Adams' result: Covering radius of subcodes*, submitted.

[26] P. SOLÉ, *Asymptotic bounds on the covering radius of binary codes*, IEEE Trans. Inform. Theory, IT-36 (1990), pp. 1470–1472.

[27] ———, *Completely regular codes and completely transitive codes*, Discrete Math., 81 (1990), pp. 193–201.

[28] P. SOLÉ AND K. G. MEHROTRA, *Generalization of the Norse bounds to codes of higher strength*, IEEE Trans. Inform. Theory, IT-37 (1991), pp. 190–192.

[29] P. SOLÉ AND T. ZASLAVSKY, *The covering radius of the cycle code of a graph*, Discrete Appl. Math., 45 (1993), pp. 63–70.

[30] D. E. TAYLOR, *Regular 2-graphs*, Proc. London Math. Soc., 35 (1977), pp. 257–274.

[31] G. TOULOUSE, *Theory of the frustration effect in spin glasses: I*, Commun. Phys., 2 (1977), pp. 115–119.

[32] A. L. WELLS, JR., *Even signings, signed switching classes, and* $(-1,1)$ *matrices*, J. Combin. Theory Ser. B, 36 (1984), pp. 194–212.

[33] T. ZASLAVSKY, *Signed graphs*, Discrete Appl. Math., 4 (1982), pp. 47–74 (Erratum, ibid., 5 (1983), p. 248).

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 554–559, November 1994

GENERALIZED HAMMING WEIGHTS OF MELAS CODES AND DUAL MELAS CODES *

G. VAN DER GEER[†] AND M. VAN DER VLUGT[‡]

Abstract. In this paper the second and the third generalized Hamming weight of Melas codes are computed and results on the second generalized Hamming weight of dual Melas codes are obtained. The authors' main tool is the theory of elliptic curves over finite fields.

Key words. generalized Hamming weights, Melas codes, dual Melas codes, elliptic curves

AMS subject classifications. primary 94B05; secondary 11T71, 14G15

1. Introduction.

Let $\mathcal{C}$ be a binary linear (n,k)-code of length n and dimension k. Since Wei proved in [W] that the performance of $\mathcal{C}$, when used on the wire-tap channel of type II, is determined by the generalized Hamming weights of $\mathcal{C}$, renewed interest in these parameters of a code has been observed (see, e.g., [C], [FTW], [GV2], [GV3], [HKY], [HTV], [W], [YKS]).

We recall that for $1 \leq r \leq k$, the rth *generalized Hamming weight* of $\mathcal{C}$ is defined by

$$d_r(\mathcal{C}) = \min\{\#S(\mathcal{D}) \,|\, \mathcal{D} \text{ is an } (n,r)\text{-subcode of } \mathcal{C}\}, \tag{1}$$

where $S(\mathcal{D})$ is the support of $\mathcal{D}$, i.e., the set of coordinate places where not all words in $\mathcal{D}$ have a coordinate 0. Since projection of $\mathcal{D}$ on a coordinate place is an $\mathbb{F}_2$-linear map, we easily see that $d_r(\mathcal{C})$ can also be defined as

$$d_r(\mathcal{C}) = \frac{1}{2^{r-1}} \min\left(\sum_{d\in\mathcal{D}} w(d) : \mathcal{D} \text{ is an } (n,r)\text{-subcode of } \mathcal{C}\right), \tag{2}$$

where $w(d)$ is the weight of the word d.

In general it is not easy to determine $d_1(\mathcal{C})$ and results on $d_r(\mathcal{C})$ for $r \geq 2$ are even more scarce until now. In this paper we obtain results on d_2 and d_3 of the classical Melas codes $M(q)$ of length $q-1$ with $q = 2^m$ and on d_2 of their dual codes. The Melas codes were introduced in [M].

2. Melas codes and their duals.

Let $\mathbb{F}_q$ be the finite field of cardinality $q = 2^m$ with $m \geq 3$. By α we denote a generator of the multiplicative group $\mathbb{F}_q^*$. The *Melas code* $M(q)$ of length $n = q-1$ is the binary cyclic code with parity check matrix

$$H = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{n-1} \\ 1 & \alpha^{-1} & \cdots & \alpha^{-(n-1)} \end{pmatrix}.$$

The code $M(q)$ is reversible (i.e., if $(c_0, c_1, \ldots, c_{n-1}) \in M(q)$, then $(c_{n-1}, \ldots, c_1, c_0) \in M(q)$). For $m \geq 3$, the dimension of $M(q)$ is $q-1-2m$ and the minimum distance

* Received by the editors January 25, 1993; accepted for publication (in revised form) September 1, 1993.

[†] Faculteit Wiskunde en Informatica, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (geer@fwi.uva.nl).

[‡] Mathematisch Instituut, Rijksuniversiteit te Leiden, Niels Bohrweg 1, 2300 RA Leiden, The Netherlands (vlugt@rulwinw.leidenuniv.nl).

is 3 for even m, while it is 5 for odd $m \geq 5$ (see [MacWS] and [SV]). In [SV] it was shown that the dual code $M(q)^{\perp}$ can be described as the binary code with words

$$c(a,b) = \left(\mathrm{Tr}\left[ax + \frac{b}{x} \right]_{x \in \mathbb{F}_q^*} \right) \quad \text{with } a, b \in \mathbb{F}_q,$$

where we denote by Tr the trace map from $\mathbb{F}_q$ to $\mathbb{F}_2$.

The weights occurring in $M(q)^{\perp}$ were determined by Lachaud and Wolfmann in [LW]. If we define $t_{\min} = \min\{t \in \mathbb{Z} : t^2 < 4q \text{ and } t \equiv 1 (\bmod 4)\}$, then the following theorem holds (see [SV]).

THEOREM 2.1. *The minimum weight in $M(q)^{\perp}$ is $(q-1+t_{\min})/2$ and the number of words of minimum weight is $(q-1)\tilde{h}(t_{\min}^2 - 4q)$.*

Here $\tilde{h}$ denotes a class number. Namely, for positive definite integral binary quadratic forms $Q(X,Y) = aX^2 + bXY + cY^2$ we have

$$\tilde{h}(\Delta) = \#\left\{ Q(X,Y) : b^2 - 4ac = \Delta \right\} / SL_2(\mathbb{Z}) - \text{equivalence}.$$

Note that $\tilde{h}(\Delta)$ is nonzero only for $\Delta \in \mathbb{Z}_{<0}$ and $\Delta \equiv 0$ or $1 (\bmod 4)$.

The theory of reduced quadratic forms yields

$$\tilde{h}(\Delta) = \#\left\{ (a,b,c) \in \mathbb{Z}^3 : b^2 - 4ac = \Delta, |b| \leq a \leq c, \text{and } b \geq 0, \text{if } a = |b| \text{ or } a = c \right\}.$$

This expression makes $\tilde{h}(\Delta)$ very easy to compute. For details we refer the reader to [O].

The multiplicative group $\mathbb{F}_q^*$ and the Galois group $\mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_2)$ of the field extension $\mathbb{F}_q/\mathbb{F}_2$ act in a canonical way on $M(q)^{\perp}$. This leads to the following lemma.

LEMMA 2.2. *If $c(a,b)$ is a word of minimum weight in $M(q)^{\perp}$, then*

(i) *for every $\beta \in \mathbb{F}_q^*$, the word $c(\beta a, \beta^{-1} b)$ has minimum weight,*

(ii) *for every $\sigma \in \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_2)$, the word $c(\sigma(a), \sigma(b))$ has minimum weight.*

Proof. (i) The map defined by $x \longmapsto \beta x$ is a permutation of $\mathbb{F}_q^*$. Hence $c(\beta a, \beta^{-1} b)$ is equivalent with $c(a,b)$ and thus has the same weight.

(ii) Likewise $\sigma \in \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_2)$ also permutes $\mathbb{F}_q^*$ and the trace map is invariant under the action of σ. This implies that $w(c(a,b)) = w(c(\sigma(a), \sigma(b)))$. □

Due to this lemma, there exist words of minimum weight of the form $c(a,1)$, necessarily with $a \in \mathbb{F}_q^*$, and their number is $\tilde{h}(t_{\min}^2 - 4q)$ (see [SV, Thm. 3.3]).

3. Determination of $d_2(M(q)^{\perp})$. First we consider the case where m is even. Then $\mathbb{F}_q$ contains a primitive third root of unity, which we denote by ρ.

THEOREM 3.1. *For even $m \geq 4$, there exists a two-dimensional subcode of $M(q)^{\perp}$ all of whose nontrivial words have minimum weight.*

Proof. Let $c(a,1)$ be a word of minimum weight. Then, according to Lemma 2.2, the weight of $c(\rho a, \rho^2)$ is also minimal. These two $\mathbb{F}_2$-independent words generate a two-dimensional subcode of the kind we are looking for, since $c(a,1) + c(\rho a, \rho^2) = c(\rho^2 a, \rho)$ also has minimum weight. □

COROLLARY 3.2. *For even $m \geq 4$, the second generalized Hamming weight of $M(q)^{\perp}$ satisfies*

$$d_2\left(M(q)^{\perp}\right) = 3d_1\left(M(q)^{\perp}\right)/2 = 3(q-1+t_{\min})/4.$$

Now we consider the case where m is odd.

THEOREM 3.3. *Let $c(a,1)$ be a word of minimum weight in $M(q)^{\perp}$. If there exists an automorphism $\sigma \in \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_2)$ such that $\mathrm{Tr}(\sigma(a)/a) = 0$, then $d_2(M(q)^{\perp}) = 3d_1(M(q)^{\perp})/2$.*

Proof. Consider the words $c(a,1)$ and $c(\beta\sigma(a),\beta^{-1})$ with $\sigma \in \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_2)$ and $\beta \in \mathbb{F}_q - \mathbb{F}_2$. Now we want the sum of these two words to be of the form $c(\gamma a,\gamma^{-1})$ for a certain $\gamma \in \mathbb{F}_q - \mathbb{F}_2$. Therefore we have to find $\sigma \in \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_2)$ and $\beta,\gamma \in \mathbb{F}_q - \mathbb{F}_2$ satisfying

$$(3)\qquad \begin{aligned} a + \beta\sigma(a) &= \gamma a, \\ 1 + \beta^{-1} &= \gamma^{-1}. \end{aligned}$$

By substitution (3) yields

$$(4)\qquad \beta^2 + \beta = a/\sigma(a)\,.$$

This equation has two solutions $\beta \in \mathbb{F}_q - \mathbb{F}_2$ if and only if $\mathrm{Tr}(a/\sigma(a)) = 0$ or $\mathrm{Tr}(\sigma^{-1}(a)/a) = 0$. Hence our condition implies the existence of a two-dimensional subcode of $M(q)^{\perp}$, all of whose nontrivial words have minimum weight. □

COROLLARY 3.4. *If $M(q)^{\perp}$ contains a minimum weight word $c(a,1)$ with $\mathrm{Tr}(a) = 0$, then $d_2(M(q)^{\perp}) = 3d_1(M(q)^{\perp})/2$.*

Due to a result in [GV1] we can translate the condition $\mathrm{Tr}(a) = 0$ into a condition that has to be satisfied by $t_{\min}$. Namely, the word $c(a,1) = (\mathrm{Tr}[ax + \frac{1}{x}]_{x\in\mathbb{F}_q^*})$ corresponds to the elliptic curve E_a with affine equation $y^2 + y = ax + \frac{1}{x}$. The $\mathbb{F}_q$-rational points on E_a form in a natural geometric way an abelian group $E_a(\mathbb{F}_q)$ of order $q + 1 - t_{\min}$ (see [S]). In [GV1] the following theorem was deduced.

THEOREM 3.5. *Let $q = 2^m$ with $m \geq 3$ and let E_a be an elliptic curve over $\mathbb{F}_q$ of the form $y^2 + y = ax + \frac{1}{x}$. Then $E_a(\mathbb{F}_q)$ has a point of order 8 if and only if $\mathrm{Tr}(a) = 0$.*

COROLLARY 3.6. *If $t_{\min} \equiv 1 (\mathrm{mod}\, 8)$, then $d_2(M(q)^{\perp}) = 3d_1(M(q)^{\perp})/2$.*

Proof. The condition $t_{\min} \equiv 1(\mathrm{mod}\, 8)$ implies that the abelian group $E_a(\mathbb{F}_q)$ on the elliptic curve involved has a subgroup of order 8. Then it follows that $E_a(\mathbb{F}_q)$ has a point of order 8 since $E_a(\mathbb{F}_q)$ has only one point of order 2 (see [S, Prop. 3.4]). We find our result if we combine Theorem 3.5 with Corollary 3.4. □

EXAMPLE 3.7. *For $m = 11$ the code $M(q)^{\perp}$ is a $(n = 2047, k = 22)$-code. From Theorem 2.1 we infer that $t_{\min} = -87 \equiv 1(\mathrm{mod}\, 8)$. Hence $d_1(M(2047)^{\perp}) = (2047 - 87)/2 = 980$ and $d_2(M(2047)^{\perp}) = 1470$.*

Remark 3.8. In the case where m is odd, there exist dual Melas codes that do not have the property $d_2 = (3/2)d_1$. Namely, for $m = 5$ we find $t_{\min} = -11$ and $\tilde{h}(t_{\min}^2 - 4q) = \tilde{h}(-7) = 1$. This implies that $c(1,1)$ has minimum weight and the set of minimum weight words is given by $\{c(\beta,\beta^{-1}) : \beta \in \mathbb{F}_q^*\}$. Then we conclude that there is no two-dimensional subcode of $M(32)^{\perp}$ containing three minimal words. In fact, $d_1(M(32)^{\perp}) = 10$ and $d_2(M(32)^{\perp}) = 16$.

Note that $m = 5$ is the only odd m for which $\tilde{h}(t_{\min}^2 - 4q) = 1$.

Finally, we derive a simple upper bound for $d_2(M(q)^{\perp})$.

THEOREM 3.9. *For odd $m \geq 7$, the second generalized Hamming weight satisfies the inequality*

$$d_2\left(M(q)^{\perp}\right) \leq (3q - 2 + 2t_{\min})/4.$$

Proof. Take a word $c(a,1)$ of minimum weight. Since $m \geq 7$ we may assume that $a \neq 1$. Then $c(a,1)$ and $c(a^2,1)$ generate a two-dimensional subcode that contains two words of weight $(q-1+t_{\min})/2$ and one word of weight $q/2$. Using (2), the upper bound easily follows. □

4. Determination of $d_2(M(q))$ and $d_3(M(q))$. In this section we label the coordinate places of a word in $M(q)$ by the elements of $\mathbb{F}_q^*$ in the same order as in the first row of the parity check matrix H for $M(q)$ given in §2.

First we consider the difficult case, i.e., the case where m is odd, $m \geq 5$. Fix two different coordinate places t_1, t_2. From (2) we see that the words of weight 5 in $M(q)$ that have ones in the places t_1, t_2 play an important role in the determination of generalized Hamming weights. If we look at the parity check matrix H it follows that these words are closely related to the solutions $(x_1, x_2, x_3) \in (\mathbb{F}_q^*)^3$ of the system of equations

$$\begin{aligned} x_1 + x_2 + x_3 &= t_1 + t_2, \\ \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} &= \frac{1}{t_1} + \frac{1}{t_2}. \end{aligned} \tag{5}$$

THEOREM 4.1. *For $q = 2^m$ with odd $m \geq 5$, the number $N_3(t_1, t_2)$ of solutions of* (5) *is at least $q - 5 - 2\sqrt{q}$.*

Proof. We replace (5) by

$$\begin{aligned} x_1 + x_2 + x_3 &= x_4, \\ \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} &= \frac{A}{x_4}, \end{aligned} \tag{6}$$

where $A = (t_1 + t_2)^2 / t_1 t_2$. Substituting the expression for x_4 in the second equation we obtain

$$(x_1 + x_2 + x_3)(x_1 x_2 + x_1 x_3 + x_2 x_3) = A x_1 x_2 x_3.$$

This equation represents a plane projective curve C_A. For odd m we have $A \neq 1$, which implies that C_A is nonsingular. Furthermore, C_A has genus 1 and an $\mathbb{F}_q$-rational point; so C_A is an elliptic curve over $\mathbb{F}_q$. The Riemann hypothesis for elliptic curves over finite fields, proved by Hasse in 1933, yields that the number $\#C_A(\mathbb{F}_q)$ of $\mathbb{F}_q$-rational points on C_A satisfies

$$\#C_A(\mathbb{F}_q) \geq q + 1 - 2\sqrt{q}.$$

Now the projective points $(x_1 : x_2 : x_3) = (1:0:0), (0:1:0), (0:0:1), (1:1:0), (1:0:1), (0:1:1)$ do not induce solutions of (6) and we conclude that (6) has at least $q - 5 - 2\sqrt{q}$ solutions for fixed $x_4 \in \mathbb{F}_q^*$. Replacing x_4 by $t_1 + t_2$ we obtain at least $q - 5 - 2\sqrt{q}$ solutions of (5). □

If we denote the number of words in $M(q)$ of weight 5 with ones in the positions t_1 and t_2 by $A_5(t_1, t_2)$, we have the following corollary.

COROLLARY 4.2. *The number of words $A_5(t_1, t_2)$ satisfies*

$$A_5(t_1, t_2) \geq (q - 5 - 2\sqrt{q})/6.$$

Proof. We verify that the solutions (x_1, x_2, x_3) of (5) satisfy $x_i \in \mathbb{F}_q^* - \{t_1, t_2\}$ for $i = 1, 2, 3$. Keeping in mind that the symmetric groups S_3 acts on the solutions of (5), Theorem 4.1 implies the inequality for $A_5(t_1, t_2)$. □

This corollary leads to the following theorem.

THEOREM 4.3. *For the Melas codes $M(q)$ of length $n = q-1$, where $q = 2^m$ with $m \geq 4$, we have for even m,*

$$\text{(i) } d_2(M(q)) = 6,$$
$$\text{(ii) } d_3(M(q)) = 9;$$

while for odd m,

$$\text{(iii) } d_2(M(q)) = 8,$$
$$\text{(iv) } d_3(M(q)) = 10.$$

Proof. (i) and (ii). For m even, $m \geq 4$, the code $M(q)$ has $(q-1)/3$ words of weight 3 (see [SV]), namely, the shifts of the word that has ones in the positions corresponding to 1, $\alpha^{(q-1)/3}, \alpha^{2(q-1)/3}$. First we conclude that $d_2(M(q)) = 6$. For a three-dimensional subcode $\mathcal{D} \subset M(q)$ with a basis of words of weight 3 we have $\#S(\mathcal{D}) = 9$. Since $M(q)$ has no words of weight 4 (see [SV]), we find that $\#S(\mathcal{D}) \geq 9$ for three-dimensional subcodes $\mathcal{D}$ that have no basis consisting of words of weight 3. Hence $d_3(M(q)) = 9$.

(iii) Since $M(q)$ has minimum distance 5, the expression (2) for $d_r(C)$ implies that $d_2(M(q)) \geq 8$. From the inequality in Corollary 4.2 we immediately see that $A_5(t_1, t_2) \geq 2$. Two words of weight 5 with ones in the positions t_1 and t_2 generate a two-dimensional subcode $\mathcal{D}$ with $S(\mathcal{D}) = 8$, which proves $d_2(M(q)) = 8$.

(iv) From (2) we derive $d_3(M(q)) \geq 10$. Let $P(t_1, t_2)$ be the set of words of weight 5 in $M(q)$ with ones in two fixed positions t_1 and t_2. Then it follows from Corollary 4.2 that the support $S(P(t_1, t_2))$ of $P(t_1, t_2)$ satisfies

$$\#S(P(t_1, t_2)) \geq 3\left(\frac{q-5-2\sqrt{q}}{6}\right) + 2.$$

Take a word ω in $P(t_1, t_2)$ with support $t_1, t_2, \omega_1, \omega_2, \omega_3$ and consider words $\neq \omega$ in $P(t_1, \omega_1)$. If the support of this set is not disjoint from $S(P(t_1, t_2))$, we have three words (namely two words, including ω, in $P(t_1, t_2)$ and a word $\neq \omega$ in $P(t_1, \omega_1)$) that generate a three-dimensional subcode of $M(q)$ that occupies 10 positions, thereby proving $d_3(M(q)) = 10$. If the supports of our second set and $P(t_1, t_2)$ are disjoint, we carry on by considering sets of words $P(t_i, \omega_j) - \{\omega\}$ for $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$. If all the supports of these sets are disjoint, we use at least

$$3\left(\frac{q-5-2\sqrt{q}}{6}\right) + 2 + 6 \cdot 3 \cdot \left(\frac{q-11-2\sqrt{q}}{6}\right) = (7q - 67 - 14\sqrt{q})/2$$

positions. Since we have only $q-1$ positions available, we get a contradiction. Hence there is a nonempty intersection, which provides us with three words, including ω, that generate a three-dimensional subcode of $M(q)$ of support size 10. This proves $d_3(M(q)) = 10$. □

For the even weight subcodes of Melas codes of length $q = 2^m$ with m odd, $m \geq 5$, it can be proved along the same lines that $d_2 = 9$ and $d_3 = 11$.

REFERENCES

[C] H. CHUNG, *The second generalized Hamming weight of double error-correcting binary* BCH *codes and their dual codes*, in Lecture Notes in Comput. Sci., Vol. 539, Springer-Verlag, New York, 1991, pp. 118–129.

[FTW] G. L. FENG, K. K. TZENG, AND V. K. WEI, *On the generalized Hamming weights of several classes of cyclic codes*, IEEE Trans. Inform. Theory, 38 (1992), pp. 1125–1130.

[GV1] G. van der Geer and M. van der Vlugt, *Kloosterman sums and the p-torsion of certain Jacobians*, Math. Ann., 290 (1991), pp. 549–563.

[GV2] ———, *On generalized Hamming weights of* BCH *codes*, IEEE Trans. Inform. Theory, to appear.

[GV3] ———, *The second generalized Hamming weight of the dual codes of double-error correcting binary* BCH *codes*, Report 92-23, University of Amsterdam, Amsterdam, The Netherlands, and Bull. London Math. Soc., to appear.

[HKY] T. Helleseth, T. Kløve, and Ø. Ytrehus, *Generalized Hamming weights of linear codes*, IEEE Trans. Inform. Theory, 38 (1992), pp. 1133–1140.

[HTV] J. W. P. Hirschfeld, M. A. Tsfasman, and S. G. Vladut, *The weight hierarchy of higher dimensional hermitian codes*, IEEE Trans. Inform. Theory, 40 (1994), pp. 275–278.

[LW] G. Lachaud and J. Wolfmann, *Sommes de Kloosterman, courbes elliptiques et codes cycliques*, Comptes Rendus Acad. Sci. Paris Série I, 305 (1987), pp. 881–883.

[M] C. M. Melas, *A cyclic code for double error correction*, IBM J. Res. Develop., 4 (1960), pp. 364–366.

[MacWS] J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*, North-Holland, Amsterdam, 1983.

[O] J. Oesterlé, *Le problème de Gauss sur le nombre de classes*, L'Enseign. Math., 34 (1988), pp. 43–67.

[S] R. Schoof, *Nonsingular plane cubic curves over finite fields*, J. Combin. Theory Ser. A, 46 (1987), pp. 183–211.

[SV] R. Schoof and M. van der Vlugt, *Hecke operators and the weight distributions of certain codes*, J. Combin. Theory Ser. A, 57 (1991), pp. 163–186.

[W] V. K. Wei, *Generalized Hamming weights for linear codes*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1412–1418.

[YKS] K. Yang, P. Kumar, and H. Stichtenoth, *On the weight hierarchy of geometric Goppa codes*, Report, University of Southern California, Los Angeles, October 1992 and IEEE Trans. Inform. Theory, to appear.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 560–570, November 1994

RAMANUJAN DIAGRAMS *

MOSHE MORGENSTERN†

Abstract. A diagram D is a graph that is of finite volume with respect to a measure (weights) on the vertices and edges. The author gives the basic definitions for a diagram, and defines the cases where it is an expander. Let Δ be the Laplacian on $L_2(D)$, and let λ be the infimum of its spectrum on the subspace of functions that are orthogonal to the constant function. The strong connection between λ being large and D being a good expander is shown. For a k-regular infinite diagram, the largest possible λ is $k - 2\sqrt{k-1}$, and when this is achieved, it is called a Ramanujan diagram. Using representation theory of PGL_2, many infinite families of infinite Ramanujan diagrams are explicitly constructed.

Key words. diagram, Ramanujan graph/diagram, expander, Laplacian, explicit constructions

AMS subject classifications. primary 05C35, secondary 05C25

1. Introduction.

An *(n,d)-expander* is a graph $\Gamma = (V,E)$ with n vertices, such that for any set $S \subseteq V$ with $|S| \leq n/2$, the set of neighbors of S, $\Gamma(S) = \{v \in V | (v,u) \in E \text{ for some } u \in S\}$, satisfies

$$|\Gamma(S) \backslash S| \geq d(1 - |S|/n)|S|.$$

Expanders have a wide area of interest, and are particularly useful in theoretical computer science when proving lower bounds, asymptotically optimal algorithms, and construction of communication networks. They were first introduced by Pinsker [Pin], Pippenger [Pip], and Margulis [Ma1]. Later, Gaber and Galil [GG] and Margulis [Ma1] gave explicit constructions for families of k regular (n,d)-expanders with fixed k, d, and $n \longrightarrow \infty$. Although Pinsker [Pin] showed that much better families exist, for a long time no one knew how to build them explicitly (even though many applications require an explicit construction) until the Ramanujan graphs were built by Lubotzky, Phillips, and Sarnak [LPS], and Margulis [Ma2], independent of one another. Ramanujan graphs are the best possible families in terms of eigenvalues, which are a measure of expansion, but better families of expander graphs may exist, for example, random graphs.

Let Γ be a finite k-regular connected graph with n vertices, and let A be its adjacency matrix. Let $\lambda(\Gamma)$ be the second smallest eigenvalue of $kI - A$. The following theorems have been proved.

THEOREM A [AM], [Ta]. *Γ is an $(n, 4\lambda/(2\lambda + k))$-expander.*

THEOREM B [Al]. *If Γ is a k-regular (n,d)-expander, then $\lambda(\Gamma) \geq d^2/4k$.*

Hence, to look for a good expander is approximately equivalent to looking for a large λ, but for this there is a limit.

THEOREM C [Al]. *Let $\{\Gamma_i = (V_i, E_i)\}_{i=1}^{\infty}$ be an infinite family of k-regular graphs. If $\lim_{i\to\infty} |V_i| = \infty$, then $\limsup_{i\to\infty} \lambda(\Gamma_i) \leq k - 2\sqrt{k-1}$.*

In [LPS], for every prime $p \neq 2$, many families of $k = p + 1$ regular graphs with $\lambda \geq k - 2\sqrt{k-1}$ are explicitly constructed. This is achieved by using the Ramanujan

* Received by the editors August 3, 1992; accepted for publication (in revised form) September 1, 1993.

† Department of Mathematics, The Hebrew University, Jerusalem, Israel. Present address, Department of Mathematics and Computer Science, Bar-Ilan University, Ramat-Gan, Israel.

conjecture (proved by Deligne [De]). These graphs are therefore called *Ramanujan graphs*. The technique is to divide the tree of $PGL_2(Q_p)$ (on which $PGL_2(Q_p)$ acts as a group of automorphisms) by a co-compact lattice that acts freely. To obtain $q+1$ regular Ramanujan graphs for q, a prime power, it is reasonable to replace Q_p by a suitable completion k of the function field $\mathbb{F}_q(x)$ over the field with q elements, see [Mo2]. Then $PGL_2(k)$ contains lattices that are not co-compact and act on the tree with "large" stabilizers (i.e., no finite index sub-lattice acts freely). The quotient is then more complicated. It is an infinite graph of finite volume by weights attached to vertices and edges (i.e., the lattice measure), and we call it a diagram. While interesting themselves, more surprising is the fact that in some, the regular part (where all weights are 1, i.e., a usual graph) contains graphs of great interest (as shown in [Mol]).

Many questions arise. What is an expanding diagram? What is the analogue of λ, and how does it relate to expansion? What is the best possible expansion? Can one reach it? In §2, we give the definitions and see that the analogue of λ (which is also called λ) satisfies Theorem A. Yet only a weaker version of Theorem B is proved: $\lambda \geq d'^2/4k$, where d' expresses the expansion of the best family of finite sub-diagrams of D that grows up to D. We leave open a few questions concerning the best possible expansion.

We then prove a stronger version of Theorem C: Even for a single k-regular infinite diagram, $\lambda \leq k - 2\sqrt{k-1}$. The same holds true for limsup λ in a family of finite diagrams where the number of vertices (and particularly where the volume) approaches infinity. The best diagrams, where $\lambda = k - 2\sqrt{k-1}$, are called Ramanujan diagrams. Finally, in §3, using representation theory of PGL_2, we explicitly construct many families of infinite, $q+1$-regular Ramanujan diagrams for every prime power q, and sketch the proof that these are Ramanujan diagrams.

2. The Laplacian of a diagram.

DEFINITION 2.1. *A* diagram *is a triple* $D = (V, E, w)$, *where* $\Gamma = (V, E)$ *is an undirected graph,* $|V| \leq \aleph_0$, $w : V \cup E \to (0, 1]$ *is the* weight function, *and for any* $e = (u, v) \in E$, $w(e) \geq w(u), w(v)$. *For* $S \subseteq V$, $\mu(S) = \sum_{u \in S} w(u)$ *is the* measure *on* D, *and we assume that* $\mu(V) < \infty$ (*we also write* $\mu(D)$ *for* $\mu(V)$). D *is called* infinite *if* $|V| = \aleph_0$. *Call* $\theta(u, v) = w(e)/w(u)$ *the* entering degree *of* $e = (u, v)$ *to* u, *and for a vertex* u, in-degree$(u) = \sum_{(u,v)\in E} \theta(u, v)$. D *is called* k-regular *if, for every* $u \in V$, in-degree$(u) = k$.

On the functions on D, an *inner product* is defined by

$$\langle f, g\rangle = \int_D f \cdot \bar{g} d\mu = \sum_{v\in V} f(v)\overline{g(v)}w(v).$$

Then $||f||$ and $L_2(D)$ are defined as usual. Define Δ, the *Laplacian* on $L_2(D)$, by

$$\begin{aligned}\Delta(f)(u) &= \text{in-degree}(u) \cdot f(u) - \sum_{(u,v)\in E} \theta(u,v) \cdot f(v) \\ &= \sum_{(u,v)\in E} \theta(u,v)(f(u) - f(v)).\end{aligned} \tag{1}$$

Using the Cauchy–Schwartz inequality, it is easy to see that if, for every $u \in V$, in-degree$(u) \leq k$, then $\Delta : L_2(D) \to L_2(D)$ is bounded and $||\Delta|| \leq 2k$.

LEMMA 2.2. *Δ is Hermitian.*

Proof.

$$\begin{aligned}
\langle \Delta f, g\rangle &= \sum_{u\in V} \Delta f(u)\overline{g(u)}w(u) = \sum_{u\in V}\sum_{(u,v)\in E} \theta(u,v)\left(f(u)-f(v)\right)\overline{g(u)}w(u)\\
&= \sum_{e=(u,v)} \left[f(u)\overline{g(u)} - f(v)\overline{g(u)} - f(u)\overline{g(v)} + f(v)\overline{g(u)}\right] w(e)\\
&= \sum_{u\in V}\sum_{(u,v)\in E} \theta(u,v)\left[\overline{g(u)-g(v)}\right] f(u)w(u)\\
&= \sum_{u\in V} \overline{\Delta g(u)} f(u)w(u) = \langle f, \Delta g\rangle . \qquad \square
\end{aligned}$$

For a k-regular connected diagram D, the constant function 1_V is an eigenvector for the eigenvalue 0. Look at the following orthogonal complement of 1_V:

$$L_2^0(D) = \left\{ f \in L_2(D) \,\middle|\, \sum_{v\in V} f(v)w(v) = 0 \right\}.$$

Because Δ is Hermitian, $L_2^0(D)$ is invariant under Δ and

$$\lambda(D) = \inf\left\{\langle \Delta f, f\rangle \,|\, f \in L_2^0(D), \|f\| = 1\right\} = \inf \operatorname{spec}\left(\Delta |\, L_2^0(D)\right).$$

DEFINITION 2.3. *A k-regular diagram D is called a $(\mu_0,\ d)$-*expander *if $\mu(V) = \mu_0$ and, for any $S \subseteq V$ with $\mu(S) \le \mu_0/2$, the set of neighbors of S, $\Gamma(S)$ satisfies $\mu(\Gamma(S)\backslash S) \ge d(1-\mu(S)/\mu_0)\mu(S)$. Let $d(D) = \sup\{d|D$ is an (μ_0, d)-expander$\}$.*

Following Alon [Al], we prove the analog of Theorem A, as follows.

THEOREM 2.4. *A connected k-regular diagram with $\mu(V) = \mu_0$ is a $(\mu_0,\ 4\lambda/(2\lambda+k))$-expander.*

Proof. For $S \subseteq V$, let $\bar{S} = V\backslash(S \cup \Gamma(S))$. Define the function $g \in L_2(D)$,

$$g(v) = \begin{cases} 1/\mu(S) & v \in S,\\ -1/\mu(\bar{S}) & v \in \bar{S},\\ \frac{1}{2}\left(1/\mu(S) - 1/\mu(\bar{S})\right) & v \in \Gamma(S)\backslash S. \end{cases}$$

Let $\sigma = \sum_{v\in V} g(v)w(v)/\mu_0$, and let $f(v) = g(v) - \sigma$. Then

$$\sum_{v\in V} f(v)w(v) = \sum_{v\in V} g(v)w(v) - \sigma \sum_{v\in V} w(v) = 0.$$

So, $f \in L_2^0(D)$, and hence $\langle \Delta f, f\rangle \ge \lambda||f||^2$. Therefore,

$$\begin{aligned}\lambda\left(\frac{1}{\mu(S)}+\frac{1}{\mu(\bar{S})}\right) &\leq \lambda\left[\left(\frac{1}{\mu(S)}-\sigma\right)^2\mu(S)+\left(\frac{1}{\mu(\bar{S})}+\sigma\right)^2\mu(\bar{S})\right]\\
&=\lambda\sum_{v\in S\cup\bar{S}} f^2(v)\,w(v)\leq\lambda\|f\|^2\leq\langle\Delta f,f\rangle\\
&=\sum_{e=(u,v)\in E}(f(u)-f(v))^2\,w(v)\\
&\leq\sum_{e\in E(S,\Gamma(S)\backslash S)}(f(u)-f(v))^2\,w(v)\\
&\quad+\sum_{e\in E(\Gamma(S)\backslash S,\bar{S})}(f(u)-f(v))^2\,w(v)\\
&\leq\frac{1}{4}\left(\frac{1}{\mu(S)}+\frac{1}{\mu(\bar{S})}\right)^2\cdot k\mu(\Gamma(S)\backslash S)\\
&\leq\frac{1}{4}\left(\frac{1}{\mu(S)}+\frac{1}{\mu(\bar{S})}\right)\cdot\frac{k\mu(\Gamma(S)\backslash S)\,\mu_0}{\mu(S)\,\mu(\bar{S})}.\end{aligned}$$

Hence,

$$\mu(\Gamma(S)\backslash S)\geq\frac{4\lambda}{k\mu_0}\mu(\bar{S})\,\mu(S)=\frac{4\lambda}{k}\left(\frac{\mu_0-\mu(S)}{\mu_0}-\frac{\mu(\Gamma(S)\backslash S)}{\mu_0}\right)\mu(S).$$

Thus,

$$\left(1+\frac{4\lambda}{k\mu_0}\mu(S)\right)\mu(\Gamma(S)\backslash S)\geq\frac{4\lambda}{k}\left(1-\frac{\mu(S)}{\mu_0}\right)\mu(S),$$

and

$$\mu(\Gamma(S)\backslash S)\geq\frac{4\lambda}{4\lambda\frac{\mu(S)}{\mu_0}+k}\left(1-\frac{\mu(S)}{\mu_0}\right)\mu(S)\geq\frac{4\lambda}{2\lambda+k}\left(1-\frac{\mu(S)}{\mu_0}\right)\mu(S).\qquad\square$$

DEFINITION 2.5. *Let D be a connected k-regular infinite diagram. A family $\mathcal{F}=\{f_i\}_{i=1}^{\infty}$ of finite sub-diagrams (which certainly are not k regular) of D is* essential *in D if $\lim_{i\to\infty}\mu(f_i)=\mu(D)$. Let*

$$d(\mathcal{F})=\inf\left\{d|\forall f\in\mathcal{F},\quad f\ is\ a\ (\mu(f),d)\text{-}expander\right\},$$

and

$$d^{\text{ess}}(D)=\sup\left\{d(\mathcal{F})\,|\mathcal{F}\ is\ essential\ in\ D\right\}$$

THEOREM 2.6. *If D is k regular, then $\lambda(D)\geq d^{\text{ess}}(D)^2/4k$.*

Of course $d^{\text{ess}}(D)\leq d(D)$, but they are equal for all diagrams that we can imagine, especially for those in §3.

Question 2.7.

(a) Is $d(D)$ equal to $d^{\text{ess}}(D)$? If not,

(b) is the last theorem true with $d(D)$, instead of $d^{\text{ess}}(D)$? If not,

(c) can we find an example where D is a good expander, $d(D)\gg d^{\text{ess}}(D)$, and $\lambda(D)$ is "small"?

If the answer to (a) and (b) is "no," then (c) seems to be a very hard task to accomplish, because the only known way to prove expansion is by showing that λ is large (or by statements implying it).

Proof of Theorem 2.6. $\lambda(D) = \inf\{\langle \Delta t, t\rangle | t \in L_2^0(D), ||t|| = 1\}$. Let $\varepsilon > 0$, and let $t \in L_2^0(D)$ be such that, $|\lambda - \langle \Delta t, t\rangle| < \varepsilon$ and $||t|| = 1$. Let $\mathcal{F}$ be an essential family in D with $|d(\mathcal{F}) - d^{\text{ess}}(D)| < \varepsilon$, and hence every $M \in \mathcal{F}$ is a $(\mu(M), d^{\text{ess}} - \varepsilon)$ expander. Let $M' \in \mathcal{F}$ be such that

$$\sum_{v \in D \setminus M'} |t(v)|^2 w(v) < \varepsilon, \tag{2}$$

and

$$\left| \sum_{v \in D \setminus M'} \Delta t(v) \cdot \overline{t(v)} \cdot w(v) \right| < \varepsilon. \tag{3}$$

Let $v_0 \in \Gamma(M') \setminus M'$, and let $M \in \mathcal{F}$ be such that $M' \cup \Gamma(M') \subseteq M$ and

$$|\sigma| \overset{\text{def}}{=} \left| \sum_{v \in D \setminus M} \frac{t(v) w(v)}{w(v_0)} \right| \leq \frac{\varepsilon}{1 + |t(v_0)| + \sum_{(u,v_0) \in E} |t(u)|}. \tag{4}$$

(It is obvious that (2) and (3) are still true for M.) The Laplacian of M, Δ_M, is defined as in (1) (here, in-degree$(u) = \sum_{(u,v) \in E(M)} \theta(u,v)$). Again, 0 is an eigenvalue of multiplicity one, and $\lambda(M) = \inf\{\langle \Delta_M f, f\rangle | f \in L_2^0(M), ||f|| = 1\}$. Define $\tilde{t} \in L_2^0(M)$ by

$$\tilde{t}(v) = \begin{cases} t(v) & v \neq v_0, \\ t(v_0) + \sum\limits_{v \in D \setminus M} \frac{t(v)w(v)}{w(v_0)} & v = v_0. \end{cases}$$

By (2) and (4),

$$\begin{aligned} \left(1 - ||\tilde{t}||^2\right) &= \left(||t||^2 - ||\tilde{t}||^2\right) \\ &\leq \left| \sum_{v \in D \setminus M} t(v) w(v) \right| + 2|t(v_0)| |\sigma| w(v_0) + \sigma^2 w(v_0) \\ &\leq \varepsilon + 2\varepsilon + \varepsilon^2 \leq 4\varepsilon. \end{aligned}$$

Furthermore, since $||\Delta_M|| \leq 2k$,

$$\left| \langle \Delta_M \tilde{t}, \tilde{t}\rangle - \frac{\langle \Delta_M \tilde{t}, \tilde{t}\rangle}{||\tilde{t}||^2} \right| \leq \left|1 - ||\tilde{t}||^2\right| \frac{\langle \Delta_M \tilde{t}, \tilde{t}\rangle}{||\tilde{t}||^2} \leq 8k\varepsilon. \tag{5}$$

Using (2) and (3), we also have that

(6)
$$
\begin{aligned}
&\left|\langle \Delta_M \tilde{t}, \tilde{t}\rangle - \langle \Delta t, t\rangle\right| \le \left|\sum_{v\in D\setminus M} \Delta t(v)\,\overline{t(v)}w(v)\right| \\
&+ \sum_{v\in M\setminus M'} \sum_{(u,v)\in E, u\notin M} \theta(v,u)\left|(t(v)-t(u))\,\overline{t(v)}\right| w(v) \\
&+ \left|\sum_{(u,v_0)\in E, u\in M} \theta(v_0,u)\left[(t(v_0)+\sigma-t(u))\left(\overline{t(v_0)+\sigma}\right) - (t(v_0)-t(u))\overline{t(v_0)}\right] w(v_0)\right| \\
&+ \left|\sum_{(u,v_0)\in E, u\in M} \theta(u,v_0)\left[(t(u)-t(v_0)-\sigma)\,\overline{t(u)} - (t(u)-t(v_0))\,\overline{t(u)}\right] w(u)\right| \\
&\le \varepsilon + 2k\varepsilon + 4k\varepsilon + k\varepsilon \le 8k\varepsilon.
\end{aligned}
$$

From (5) and (6) we have that

$$\lambda = \lambda(D) \ge \langle \Delta t, t\rangle - \varepsilon \ge \frac{\langle \Delta_M \tilde{t}, \tilde{t}\rangle}{\|\tilde{t}\|^2} - 17k\varepsilon \ge \lambda(M) - 17k\varepsilon. \tag{7}$$

Following an idea of Dodziuk [Do], we now estimate $\lambda(M)$. Let $f \in L_2^0(M)$ be an eigenfunction for $\lambda(M)$ with $\|f\| = 1$. Since Δ_M is Hermitian, we may assume that f is real. Let $M^+ = \{v \in M | f(v) > 0\}$, and

$$g(v) = \begin{cases} f(v) & v \in M^+, \\ 0 & \text{otherwise.} \end{cases}$$

Since f can be replaced by $(-f)$, we may assume that $\mu(M^+) \le \mu(M)/2$. Let us agree that when we write $\sum_{e=(u,v)\in E}$, we sum once for each edge (the direction is of no consequence), but $\sum_v \sum_{e=(v,u)\in E}$ implies that we sum each edge twice, once for each direction.

Observe that

$$
\begin{aligned}
&\sum_{e=(u,v)\in E(M)} w(e)\,[g(v)-g(u)]^2 \\
&\quad= \sum_{v\in M} \sum_{e=(v,u)\in E(M)} w(e)\,[g(v)-g(u)]\,g(v) \\
&\quad= \sum_{v\in M} \sum_{e=(v,u)\in E(M)} \theta(v,u)\,[g(v)-g(u)]\,g(v)\,w(v) = \langle \Delta g, g\rangle \\
&\quad= \sum_{v\in M^+} \sum_{(v,u)\in E(M)} \theta(v,u)\,[f(v)-g(u)]\,f(v)\,w(v) \\
&\quad\le \sum_{v\in M^+} \sum_{(v,u)\in E(M)} \theta(v,u)\,[f(v)-f(u)]\,f(v)\,w(v) \\
&\quad= \sum_{v\in M^+} \Delta f(v)\cdot f(v)\,w(v) = \lambda(M) \sum_{v\in M^+} |f(v)|^2 w(v) = \lambda(M)\,\|g\|^2 .
\end{aligned} \tag{8}
$$

The main idea behind obtaining the inequality above is that

$$\sum_{v\in M^+}\ \sum_{(u,v)\in E(M),u\notin M^+} \theta(v,u)\, f(u)\, f(v)\, w(v) \le 0.$$

Consider now the sum

$$\sum = \sum_{e=(u,v)\in E(M)} \left|g^2(u) - g^2(v)\right| w(e).$$

Using the Cauchy–Schwartz inequality and (8), we have that

(9)

$$\begin{aligned}
\sum &= \sum_{e=(u,v)\in E(M)} (g(u)-g(v))\,(g(u)+g(v))\, w(e) \\
&\le \left[\sum_{e=(u,v)\in E(M)} (g(u)-g(v))^2 w(e)\right]^{1/2} \left[\sum_{e=(u,v)\in E(M)} (g(u)-g(v))^2 w(e)\right]^{1/2} \\
&\le \lambda^{1/2}(M)\,\|g\| \left[\sum_{e=(u,v)\in E(M)} \left(g^2(u)+g^2(v)\right) w(e)\right]^{1/2} \\
&= \lambda^{1/2}(M)\,\|g\| \left[\sum_{v\in M} g^2(v) \sum_{e=(v,u)\in E(M)} \theta(v,u)\, w(v)\right]^{1/2} \\
&= \lambda^{1/2}(M)\,\|g\|\, k^{1/2}\,\|g\| = (k\lambda(M))^{1/2}\,\|g\|^2.
\end{aligned}$$

From the other side, let $0 = s_r < \cdots < s_1$ be all the different values of g^2, and let $L_i = \{v \in M | g^2(v) \ge s_i\}.M = L_r \supset L_{r-1} = M^+ \supset \cdots \supset L_1$. It is easy to see that

$$\begin{aligned}
\sum &= \sum_{i=1}^{r}\ \sum_{\substack{v\in M\\ g^2(v)=s_i}}\ \sum_{j>i}\ \sum_{\substack{e=(v,u)\in E(M)\\ g^2(u)=s_j}}\ \sum_{k=i}^{j-1} (s_k - s_{k+1})\, w(e) \\
&= \sum_{i=1}^{r-1} (s_i - s_{i+1}) \sum_{e=(v,u),v\in L_i,u\notin L_i} w(e) \\
&\ge \sum_{i=1}^{r-1} (s_i - s_{i+1}) \sum_{e=(v,u),v\in L_i,u\notin L_i} w(u) \ge \sum_{i=1}^{r-1} (s_i - s_{i+1})\, \mu(\Gamma(L_i)\backslash L_i).
\end{aligned}$$

Recall that M is a $(\mu(M), d^{\text{ess}}(D)-\varepsilon)$ expander, and that for $i \le r-1, L_i \subseteq M^+$. Hence $\mu(L_i) \le \frac{1}{2}\mu(M)$. We obtain that

$$\begin{aligned}
\sum &\ge \sum_{i=1}^{r-1} (s_i - s_{i+1})\left(d^{\text{ess}} - \varepsilon\right) \frac{1}{2}\mu(L_i) \\
\text{(10)}\qquad &= \frac{\left(d^{\text{ess}}(D)-\varepsilon\right)}{2} \left[s_1\mu(L_i) + \sum_{i=2}^{r-1} s_1\left(\mu(L_i) - \mu(L_{i-1})\right)\right] \\
&= \frac{\left(d_{\text{ess}}(D)-\varepsilon\right)}{2}\,\|g\|^2.
\end{aligned}$$

Putting together (9) and (10), we have that

$$\lambda (M) \geq \frac{(d^{\mathrm{ess}} (D) - \varepsilon)^2}{4k}. \tag{11}$$

This, together with (7), gives that for any $\varepsilon > 0, \lambda(D) \geq ((d^{\mathrm{ess}}(D) - \varepsilon)^2/4k) - 17\varepsilon$, and hence $\lambda(D) \geq (d^{\mathrm{ess}})^2/4k$. □

Although we cannot say what is the best possible expansion, we can say what is the best possible λ.

THEOREM 2.8. (a) *For a k-regular infinite diagram $D, \lambda(D) \leq k - 2\sqrt{k-1}$.* (b) *For a family $\{D_i = (V_i, E_i, w_i)\}_{i=1}^{\infty}$ of finite k-regular diagrams subject to $|V_i| \to \infty$ (and especially if $\mu(V_i) \to \infty$), $\limsup_{i\to\infty} \lambda(D_i) \leq k - 2\sqrt{k-1}$.*

Proof. We simply modify the proof of Alon and Boppana (see [Al]) to suit the case of diagrams. Let $n \in \mathbb{N}$, and assume that (u_1, v_1) and (u_2, v_2) are two edges of D of distance greater than $2n + 2$. Let

$$V_0^{(1)} = \{u_1, v_1\},$$
$$V_i^{(1)} = \left\{v \in V | \text{distance} \left(v, V_0^{(1)}\right) = i\right\} \quad i = 1, 2, 3, \ldots,$$

and

$$V_0^{(2)}, V_1^{(2)}, \ldots,$$

are defined similarly. Define a function $f : V \to \mathbb{R}$ as follows:

$$f(v) = \begin{cases} a_1 (k-1)^{-i/2} & v \in V_i^{(1)} \text{ and } i \leq n, \\ a_2 (k-1)^{-i/2} & v \in V_i^{(2)} \text{ and } i \leq n, \\ 0 & \text{otherwise,} \end{cases}$$

where a_1, a_2 are chosen so that $f \in L_2^0(D)$. Clearly, $||f||^2 = A_1 + A_2$, where $A_j = a_j^2 \sum_{i=0}^n \mu(V_i^{(j)})(k-1)^{-i}, j = 1, 2$. From the other side, let $E(V_i^{(j)}, V_{i+1}^{(j)}) = \{e = (u, v) | u \in V_i^{(j)}, v \in V_{i+1}^{(j)}\}$. Then

$$\langle \Delta f, f \rangle = \sum_{e=(u,v)\in E} w(e) [f(u) - f(v)]^2 = B_1 + B_2,$$

where

$$\begin{aligned} B_j &= a_j^2 \sum_{i=0}^{n-1} \mu\left(E\left(V_i^{(j)}, V_{i+1}^{(j)}\right)\right) \left(\frac{1}{(k-1)^{i/2}} - \frac{1}{(k-1)^{(i+1)/2}}\right)^2 \\ &\quad + \mu\left(E\left(V_n^{(j)}, V_{n+1}^{(j)}\right)\right) \left(\frac{1}{(k-1)^{n/2}}\right)^2 \\ &\leq a_j^2 \left[\sum_{i=0}^{n-1} (k-1) \mu\left(V_i^{(j)}\right) \left(\frac{1}{(k-1)^{i/2}} - \frac{1}{(k-1)^{(i+1)/2}}\right)^2 \right. \\ &\quad \left. + (k-1) \mu\left(V_n^{(j)}\right) \frac{1}{(k-1)^n}\right] \\ &= a_j^2 \left[\sum_{i=0}^{n} \frac{\mu\left(V_i^{(j)}\right)}{(k-1)^i} \left(k - 2\sqrt{k-1}\right) + \left(2\sqrt{k-1} - 1\right) \frac{\mu\left(V_n^{(j)}\right)}{(k-1)^n}\right] \\ &\leq A_j \left[\left(k - 2\sqrt{k-1}\right) + \frac{2\sqrt{k-1} - 1}{n+1}\right]. \end{aligned}$$

For the two inequalities above, observe that, if $u \in V_i^{(j)}$ and $e_1 = (u, v_1), \ldots, e_s = (u, v_s)$ are all edges that touch u, then at most $s-1$ of $\{v_1, \ldots, v_s\}$ are in $V_{i+1}^{(j)}$. Let us assume that v_1 is not. Since $\theta(u, v_r) \geq 1$ for $1 \leq r \leq s$, we have that

$$\begin{aligned}\sum_{r=2}^{s} w(v_r) &\leq \sum_{r=2}^{s} w(e_r) = w(u) \sum_{r=2}^{s} \theta(u, v_r) \\ &\leq w(u)\left[\sum_{r=1}^{s} \theta(u, v_r) - 1\right] = (k-1)\, w(u).\end{aligned} \tag{12}$$

From this, the first inequality is clear. For the second inequality, it is enough to show that $\mu(V_{i+1}^{(j)}) \leq (k-1)\mu(V_i^{(j)})$, which is again clear from (12).

Combining the above, we have that

$$\lambda(D) \leq \frac{\langle \Delta f, f\rangle}{\langle f, f\rangle} = \frac{B_1 + B_2}{A_1 + A_2} \leq \max\left\{\frac{B_1}{A_1}, \frac{B_2}{A_2}\right\} \leq k - 2\sqrt{k-1} + \frac{2\sqrt{k-1} - 1}{n+1}.$$

Because in both cases of our theorem we are able to do it for n as large as we want, we obtain $\lambda(D) \leq k - 2\sqrt{k-1}$ for an infinite diagram, and $\limsup \lambda(D_i) \leq k - 2\sqrt{k-1}$ in the other case. □

DEFINITION 2.9. (a) *An infinite k-regular diagram with* $\lambda(D) = k - 2\sqrt{k-1}$ *is called a* Ramanujan diagram.

(b) *A family* $\{D_i = (V_i, E_i, w_i)\}_{i=1}^{\infty}$ *of finite k-regular diagrams, subject to* $|V_i| \to \infty$ *and* $\lim_{i\to\infty} \lambda(D_i) = k - 2\sqrt{k-1}$, *is called a family of* Ramanujan diagrams.

3. Explicit construction of infinite Ramanujan diagrams. $\mathbb{F}_q$ is the field with q elements, v_∞ is the valuation at infinity of $\mathbb{F}_q(x)$ (i.e., $v_\infty(f(x)/g(x)) =$ degree(g) − degree(f)). It is well known that $k = \mathbb{F}_q((1/x))$—the Laurent series in the variable $1/x$ over $\mathbb{F}_q$—is the completion of $\mathbb{F}_q(x)$ with respect to the metric $|a| = q^{-v_\infty(a)}$, and that $\mathcal{O} = F[[1/x]]$—the Taylor series in $1/x$ over $\mathbb{F}_q$—are the integers of k. Let $G = PGL_2(k)$ be the 2×2 invertible matrices over k, divided by its center. $K = PGL_2(\mathcal{O})$ (i.e., 2×2 matrices over $\mathcal{O}$ with determinant in $\mathcal{O}^*$, divided by its center) is a maximal compact subgroup. In [Se] the structure of the $q+1$ regular tree is put on G/K, and it is shown that $\Gamma(1) = PGL_2(\mathbb{F}_q[x])$ is a lattice in G that is not co-compact (i.e., discrete and of finite co-volume, but $\Gamma(1)\backslash G$, and therefore $\Gamma(1)\backslash G/K$, are infinite). Since for every $g(x) \in \mathbb{F}_q[x]$,

$$\Gamma(g) = \left\{\begin{pmatrix} a & b \\ c & d\end{pmatrix} \in \Gamma(1) \,\middle|\, a-1, d-1, b, c, \equiv 0\,(\operatorname{mod} g(x))\right\}$$

is of finite index in $\Gamma(1)$, it is also such a lattice in G. $\Gamma(g)$ acts (by multiplications from the left) as a group of automorphisms of the tree G/K. The quotient $D_g = \Gamma(g)\backslash G/K$ is an infinite $q+1$ regular diagram (as a quotient of that $q+1$ regular tree), under the weights $w(\Gamma(g)aK) = |\{\gamma \in \Gamma(g) | \gamma \cdot aK = aK\}|^{-1}$, and in the same way for an edge $e, w(e) = |\{\gamma \in \Gamma(g) | \gamma e = e\}|^{-1}$. See [Mo1] for the exact structure of D_g. In [Mo1] we also proved the following theorem.

THEOREM 3.1. *D_g is a Ramanujan diagram.*

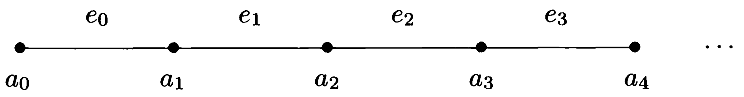
Main steps of the proof. The continuous spectrum of the Laplacian Δ on $L_2^0(D_g)$ is well known from the theory of Eisenstein series ([Ge, §8], for example). It is exactly the segment $[k-2q^{1/2}, k+2q^{1/2}]$ (see also a direct computation for the case of $\Gamma(1)$ in [Ef]).

As in [Lu], it is shown that λ is an eigenvalue of Δ if and only if the irreducible, unitary, class one representation ρ^λ appears in the regular representation of G in $L_2(\Gamma(g)\backslash G)$, where ρ^λ is such that its unique spherical function is an eigenvector for the Laplacian on $L_2(\Gamma(g)\backslash G)$ with eigenvalue λ. Then, using strong approximation, ρ^λ is lifted to a cuspidal, class-one representation of the Adeles of G. So, by Drinfeld's theorem [Dr], it is a principle series representation. In particular, $k - 2q^{1/2} \leq \lambda \leq k + 2q^{1/2}$. □

Although we give D_g as a quotient of the infinite tree, it can be made very explicit. In [Mol] we showed that D_g is the following diagram: Let degree(g) = $d, \mathcal{R} = \mathbb{F}_q[x]/g(x)\mathbb{F}_q[x]$ be represented by all polynomials (over $\mathbb{F}_q$) of degree smaller than d, (if g is irreducible this is just $\mathbb{F}_{q^d}$). $H = PGL_2(\mathcal{R}), B_0 = PGL_2(\mathbb{F}_q)$, and

$$B_n = \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} | a \in \mathbb{F}_q^*, b \in \mathcal{R}, \text{degree}\,(b) \leq n \right\} \quad n = 1, 2, \dots, d-1.$$

The vertices of D_g are divided into sets $V_0, V_1, \dots, V_{d-1}$, and cusps. Where, for $i = 0, 1, 2, \dots, d-2, V_i = H/B_i$, edges are allowed only between V_i and V_{i+1} and an edge exists between aB_i and bB_{i+1}, if and only if $aB_i \cap bB_{i+1} \neq \phi$. So the edges between aB_i and bB_{i+1} are just the cosets of $B_i \cap B_{i+1}$, which are contained in $aB_i \cap bB_{i+1}$ (if there are such cosets). Everything is finite and all weights are 1. Then on every vertex a_0 of V_{d-1}, the same infinite cusps of the form



are glued, where $w(a_i) = w(e_i) = q^{-i}$. This gives D_g explicitly, and one can easily see that D_g is trivial almost everywhere, except for the graph between V_0 and V_1, which is of great interest. For $q \geq 5$, it results in a bounded concentrator. (For more details see [Mol].)

REFERENCES

[Al] N. Alon, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83–96.

[AM] N. Alon and V. Milman, λ_1, *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[De] P. Deligne, *Forms modulaires et representations*, 1-*adiques*, Sem. Bourbaki exp. 355, Lecture Notes in Mathematics, Vol. 179, Springer-Verlag, Berlin, 1968/69.

[Do] J. Dodziuk, *Difference equation, isoperimetric inequality and transcience of certain random walks*, Trans. Amer. Math. Soc., 284 (1984), pp. 787–794.

[Dr] V. G. Drinfeld, *The proof of Peterson's Conjecture for $GL(2)$ over global field of characteristic p*, Functional Anal. Appl., 22 (1988), pp. 28–43.

[Ef] I. Efrat, *Automorphic spectra on the tree of PGL_2*, Enseign. Math. (2), 37 (1991), pp. 31–34.

[GG] O. Gaber and Z. Galil, *Explicit construction of linear sized superconcentrators*, J. Comput. System Sci., 22 (1981), pp. 407–420.

[Ge] S. Gelbart, *Automorphic Form on Adele Groups*, Princeton University Press, Princeton, NJ, 1975.

[LPS] A. Lubotzky, R. Philips, and P. Sarnak, *Ramanujan graphs*, Combinatorica, 8 (1988), pp. 261–277.

[Lu] A. Lubotzky, *Discrete Groups, Expanding Graphs and Invariant Measures*, Progress in Mathematics, Birkhauser-Verlag, Basel, Switzerland, to appear.

[Mal] G. A. Margulis, *Explicit construction of concentrators*, Problems Inform. Transmission, 10 (1975), pp. 325–332.

[Ma2] G. A. MARGULIS, *Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and superconcentrators*, Problemy Peredachi Informatsii, 24 (1988), pp. 51–60. (In Russian.) English translation in Problems Inform. Transmission, 24 (1988), pp. 39–46.

[Mo1] M. MORGENSTERN, *Natural bounded concentrators*, Combinatorica, to appear.

[Mo2] ———, *Existence and explicit construction of* $q+1$ *regular Ramanujan graphs for every prime power* q, J. Combin. Theory Ser. B to appear.

[Pin] M. S. PINSKER, *On the complexity of a concentrator*, in Proc. 7th International Teletraffic Conf., Stockholm, 1973.

[Pip] N. PIPPENGER, *Superconcentrators*, SIAM J. Comput., 6 (1972), pp. 298–304.

[Se] J. P. SERRE, *Trees*, Springer-Verlag, Berlin, New York, 1980.

[Ta] R. M. TANNER, *Explicit concentrators from generalized* N-gons, SIAM J. Algebraic Discrete Methods, 5 (1984), pp. 287–294.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 571–584, November 1994

005

ON PATH-TOUGH GRAPHS *

PETER DANKELMANN[†], THOMAS NIESSEN[†], AND INGO SCHIERMEYER[‡]

Abstract. A graph G is called path-tough, if, for each nonempty set S of vertices, the graph G–S can be covered by at most $|S|$ vertex disjoint paths. The authors prove that every graph of order n and minimum degree at least $[3/(6+\sqrt{3})]n$ is Hamiltonian if and only if it is path-tough. Similar results involving the degree sum of two or three independent vertices, respectively, are given. Moreover, it is shown that every path-tough graph without three independent vertices of degree 2 contains a 2-factor. The authors also consider complexity aspects and prove that the decision problem of whether a given graph is path-tough is NP-complete.

Key words. Hamilton cycle, path-tough graph, toughness, 2-factor, degree condition, NP-completeness

AMS subject classifications. 05C38, 05C45, 68C25

1. Introduction.

It is well known that 1-toughness is a necessary condition for a graph to be Hamiltonian, but it is not sufficient. The purpose of our work is the study of some properties of so-called *path-tough* graphs, where path-toughness is another necessary condition for Hamiltonicity related to 1-toughness.

We begin with some definitions and convenient notation. A good reference for any undefined terms is [8]. All graphs here are finite and undirected without loops or multiple edges. The vertex set of a graph G is $V(G)$, and the edge set of G is $E(G)$. We use $\omega(G)$ for the number of components of G, and $\alpha(G)$ denotes the cardinality of a maximum set of independent vertices. If $v \in V(G)$, then $N_G(v)$ is the set of all vertices in $V(G)$ adjacent to v and $d_G(v) = |N_G(v)|$ is the degree of v. The minimum degree and the maximum degree of G are denoted by $\delta(G)$ and $\Delta(G)$, respectively. Furthermore, for $k \geq 2$, let $\sigma_k(G) = \min \sum_{i=1}^{k} d_G(v_i)$, where the minimum is taken over all independent subsets $\{v_1, v_2, \dots, v_k\} \subseteq V(G)$. Here we adopt the convention $\min \varnothing = +\infty$, and so $\sigma_k(G) = +\infty$, if $\alpha(G) < k$. For disjoint $U, V \subseteq V(G)$, we denote by $G[U]$ the subgraph of G induced by U, and by $G[U,V]$ the bipartite subgraph of G having vertex set $U \cup V$ and edge set $\{uv \in E(G) | u \in U, v \in V\}$. Moreover, we let $e(U,V) = |E(G[U,V])|$. As introduced by Chvátal [12], a graph G is t-tough, if $|S| \geq t\omega(G-S)$ for every subset $S \subseteq V(G)$ with $\omega(G-S) > 1$. The *toughness* of G, denoted by $t(G)$, is the maximum value of t for which G is t-tough ($t(K_n) = +\infty$ for $n \geq 1$). Finally, the *path-covering number* $\mu(G)$ of G is the minimum number of pairwise vertex-disjoint paths that cover all vertices of G. This parameter has been investigated by Goodman and Hedetniemi [18] and Boesch, Chen, and McHugh [7]. If no ambiguity arises, we sometimes write $d(v)$ instead of $d_G(v), \alpha$ instead of $\alpha(G)$, and so forth.

Let G be a graph with a Hamilton cycle C and let S be a nonempty subset of $V(G)$. Obviously, we have $\mu(G-S) \leq \mu(C-S) \leq |S|$, and therefore $\mu(G-S) \leq |S|$ for

* Received by the editors October 28, 1992; accepted for publication (in revised form) September 1, 1993.

[†] Lehrstuhl II fur Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, Templergraben 55, 52056 Aachen, Germany.

[‡] Lehrstuhl C fur Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, Templergraben 55, 52056 Aachen, Germany.

all nonempty sets $S \subseteq V(G)$ is a necessary condition for Hamiltonicity. This condition appeared in Häggkvist [19] and Hendry [21] (see §§2 and 5), and, following a suggestion of J. A. Bondy (personal communication), we call graphs with this property *path-tough.* The reason for this notion inspired our first proposition, which follows easily from the inequality $\mu(G) \geq \omega(G)$.

PROPOSITION 1. *Let G be a path-tough graph. Then G is 1-tough or $G \cong \bar{K}_2$.*

For any path-tough graph G, it follows directly from the definition that every vertex-deleted subgraph $G-v, v \in V(G)$, is traceable (that is, there exists a Hamilton path). The next proposition shows that the converse also holds.

PROPOSITION 2. *A graph G is path-tough if and only if $G-v$ is traceable for every vertex $v \in V(G)$.*

Proof. We consider only the nontrivial implication. Let S be any nonempty subset of $V(G)$. Choose a vertex $v \in S$ and let $S' = S - \{v\}$. Since $G - v$ has a Hamilton path P, we have

$$\mu(G-S) \leq \mu(P-S') \leq |S'|+1 = |S|,$$

as required. □

2. Main results.

Two well-known sufficient conditions for Hamiltonicity are due to Dirac [13] and Ore [24].

THEOREM 1. (Dirac [13]). *Let G be a graph on $n \geq 3$ vertices. If $\delta(G) \geq n/2$, then G is Hamiltonian.*

THEOREM 2 (ORE [24]). *Let G be a graph on $n \geq 3$ vertices. If $\sigma_2(G) \geq n$, then G is Hamiltonian.*

The complete bipartite graph $K_{k,k-1}$, which is not 1-tough, shows that the degree bounds for σ_2 and δ are the best possible. Jung [23] improved Theorem 2 for 1-tough graphs as follows.

THEOREM 3. *Let G be a 1-tough graph with $n \geq 11$ vertices. If $\sigma_2 \geq n-4$, then G is Hamiltonian.*

COROLLARY 1. *Let G be a 1-tough graph with $n \geq 11$ vertices. If $\delta \geq \frac{1}{2}(n-4)$, then G is Hamiltonian.*

Let the *net* N be the graph obtained by adding three independent edges between a triangle and three independent vertices. For $n \geq 7$, odd, the graph $J_n = K_{(n-5)/2} + [(n-7)/2]K_1 \cup N)$, which is not path-tough, shows that the degree bounds for σ_2 and δ are the best possible (here "+" denotes the join, and "∪" denotes the union of two disjoint graphs).

We now present four degree conditions for path-tough graphs to be Hamiltonian. The proofs are given in §4.

THEOREM 4. *Let G be a path-tough graph with $n \geq 3$ vertices. If $\delta \geq [3/(6+\sqrt{3})]n$, then G is Hamiltonian.*

This theorem improves the first part of the following result due to Häggkvist [19].

THEOREM 5. *Let G be a non-Hamiltonian graph of order n with $\delta \geq \frac{8}{17}(n-1)$. Then G is not path-tough. Moreover, G contains a set S of at least $3\delta - n + 2$ vertices such that $G-S$ cannot be covered by $|S|$ vertex disjoint paths.*

Let p, q, r be integers such that $p \geq q \geq r \geq 3$ and $p+q+r=n$. Let G_{pqr} denote the graph with n vertices obtained from three disjoint complete graphs $H_1 = K_p, H_2 = K_q$, and $H_3 = K_r$ by adding the edges of two triangles between two disjoint triples of vertices, each containing one vertex of each of H_1, H_2, H_3. It is easy to confirm that G_{pqr} is an edge-maximal non-Hamiltonian path-tough graph. G_{pqr} has minimum degree $\delta = r-1$, and so, for $p=q=r$, we have $\delta = (n-3)/3$, showing that the bound in Theorem 4 must be at least $(n-2)/3$. We conjecture that Theorem

4 remains valid under this weaker hypothesis, with the exception of the Petersen graph, which is a non-Hamiltonian, path-tough graph with $\delta = (n-1)/3$. A stronger conjecture has already been posed by the third author, Schiermeyer, at the 13th British Combinatorial Conference, 1991.

CONJECTURE 1. *Let G be a path-tough graph with $n \geq 3$ vertices. If $\sigma_3 \geq n-2$, then G is Hamiltonian or isomorphic to the Petersen graph.*

The proof of Theorem 4 is based on our following result.

THEOREM 6. *Let G be a graph. If $\delta + \Delta \geq n \geq 3$, then either G is Hamiltonian or G is not path-tough.*

For $p = n-6$ and $q = r = 3$, the graph G_{pqr} has $\delta + \Delta = n-3$, showing that the bound in Theorem 6 is almost the best possible.

Replacing "path-tough" by "1-tough" in Theorem 6, we find that the lower bound for $\delta + \Delta$ must be at least $(3n-6)/2$, as indicated by the graph J_n, for $n \geq 9$.

THEOREM 7. *Let G be a path-tough graph with $n \geq 3$ vertices. If $\sigma_2 \geq \frac{4}{5}(n-1)$, then G is Hamiltonian.*

Again, the graph G_{pqr} for $p = q = r$ shows that the bound for σ_2 must be at least $(2n-4)/3$, and the Petersen graph has $\sigma_2 = (2n-2)/3$. We also refer to Conjecture 1.

The proof of Theorem 7 is based on our following result, which is an extension of Theorem 6.

THEOREM 8. *Let G be a 2-connected graph. If $\frac{1}{2}\sigma_2 + \Delta \geq n \geq 3$, then G is either Hamiltonian or not path-tough.*

As in the case of Theorem 6, replacing "path-tough" by "1-tough" in Theorem 8, we find that the lower bound for $\frac{1}{2}\sigma_2 + \Delta$ must be at least $(3n-6)/2$, as indicated by the graph J_n, for $n \geq 11$.

Bondy [9] extended Theorems 1 and 2 to independent sets with more than two vertices as follows.

THEOREM 9. *Let G be a k-connected graph. If $\sigma_{k+1} \geq \frac{1}{2}((k+1)(n-1)+1)$, then G is Hamiltonian.*

For 1-tough graphs, Faßbender [16] proved the following extension of Theorem 9 in the case where $k = 2$.

THEOREM 10. *Let G be a 1-tough graph with $n \geq 13$ vertices. If $\sigma_3 \geq \frac{1}{2}(3n-14)$, then G is Hamiltonian.*

For $n \geq 13$, the graph J_n has $\sigma_3 = \frac{1}{2}(3n-15)$, showing that the bound in Theorem 10 is the best possible. For path-tough graphs, we obtain the following result.

THEOREM 11. *Let G be a path-tough graph on $n \geq 3$ vertices. If $\sigma_3 \geq [(9+\sqrt{3})/(6+\sqrt{3})]n$, then G is Hamiltonian.*

Proof. If $\delta \geq [3/(6+\sqrt{3})]n$, then G is Hamiltonian by Theorem 4. If $\delta < [3/(6+\sqrt{3})]n$, then $\sigma_3 > n + \delta$, and so G is Hamiltonian by the following result of Bauer et al. [3]. □

THEOREM 12 (see [3]). *Let G be a 2-connected graph. If $\sigma_3 \geq n + \delta$, then G is Hamiltonian.*

The graph G_{pqr} for $p = q = r$ shows that the lower bound for σ_3 must be at least $n-2$, and the Petersen has $\sigma_3 = n-1$. Again, we refer to Conjecture 1.

Actually, it is not possible to apply the proof concept of Theorems 6 and 8 to obtain a stronger σ_3-condition. The reason is that the corresponding condition $\frac{1}{3}\sigma_3 + \Delta \geq n$ does not hold, since the graph G_{pqr} with $p = n-6$ and $q = r = 3$ has $\frac{1}{3}\sigma_3 + \Delta = (4n-18)/3$.

We now turn to the existence of k-regular spanning subgraphs, called k-factors. Since every Hamiltonian graph has a 2-factor, the case where $k = 2$ can be seen as a relaxation of the Hamilton cycle problem. Proving a conjecture of Chvátal [12], Enomoto et al. [15] established the next theorem.

THEOREM 13. *Let k be a positive integer and let G be a k-tough graph with n vertices. If $n \geq k+1$ and kn even, then G has a k-factor.*

Furthermore, they showed that this result is the best possible in the following sense. Let k be a positive integer and let $\varepsilon > 0$. Then there exists a $(k-\varepsilon)$-tough graph G with n vertices, where $n \geq k+1$ and kn even, which has no k-factor.

In the case where $k = 2$, we prove that "2-tough" can be replaced by "path-tough."

THEOREM 14. *Let G be a path-tough graph of order $n \geq 3$. If $\sigma_3 \geq 7$, then G has a 2-factor.*

Note that $t(G) \geq 2$ implies that $\delta(G) \geq 4$, and therefore $\sigma_3 \geq 12$; so, compared with Theorem 13, the degree condition is no further restriction. Moreover, it is impossible to omit the condition $\sigma_3 \geq 7$. To see this, we construct path-tough graphs without 2-factors as follows. For integers $p \geq q \geq 3$, let the vertex set of the graph H_{pq} consist of the disjoint sets $\{v_1, \ldots, v_p\}, \{u_1, \ldots, u_q\}$, and $\{w_1, w_2, w_3\}$, and let $E(H_{pq}) = \{v_iv_j | 1 \leq i < j \leq p\} \cup \{u_iu_j | 1 \leq i < j \leq q\} \cup \{v_iw_i | i = 1,2,3\} \cup \{u_iw_i | i = 1,2,3\}$.

We conclude this section with some complexity results. Here we use the terminology of Garey and Johnson [17]. It is well known that HAMILTON CYCLE and HAMILTON PATH are NP-complete problems. Bauer, Hakimi, and Schmeichel [4] proved that NOT 1-TOUGH, the decision problem of whether a given graph is not 1-tough, is NP-complete. Let PATH-TOUGH denote the decision problem of whether a given graph G is path-tough. As for the problems mentioned above, it is easy to see that PATH-TOUGH is a member of NP (use Proposition 2). Transforming HAMILTON PATH to PATH-TOUGH, we prove the following results.

THEOREM 15. *PATH-TOUGH is* NP*-complete.*

We now consider HAMILTON CYCLE and PATH-TOUGH, restricted to graphs with $\delta \geq (\frac{1}{2} - \varepsilon)n$, where $\varepsilon > 0$ is fixed and n denotes the order of the graph. These restrictions are respectively denoted by HAMILTON CYCLE $(\delta \geq (\frac{1}{2} - \varepsilon)n)$ and PATH-TOUGH $(\delta \geq (\frac{1}{2} - \varepsilon)n)$.

THEOREM 16. *HAMILTON CYCLE $(\delta \geq (\frac{1}{2} - \varepsilon)n)$ is* NP*-complete for every $\varepsilon > 0$.*

Note that HAMILTON CYCLE restricted to graphs with $\delta \geq \frac{1}{2}n$ is "easy" by Theorem 1.

By Theorem 4, we immediately obtain Theorem 17.

THEOREM 17. *PATH-TOUGH $(\delta \geq (\frac{1}{2} - \varepsilon)n)$ is* NP*-complete for every $\varepsilon > 0$.*

3. Preliminary results.

In this section, we state some results that we use in the proofs of our main results in the next section.

We start with the following well-known theorem due to Ore [24].

THEOREM 18. *Let G be a graph with n vertices and let u, v be a pair of nonadjacent vertices. If*

$$d_G(u) + d_G(v) \geq n, \tag{1}$$

then G is Hamiltonian if and only if $G + uv$ is Hamiltonian.

For a pair u, v of nonadjacent vertices of a graph G, we define $\lambda_{uv} = |N_G(u) \cap N_G(v)|$, $T_{uv} = \{w \in V(G) - \{u,v\} | u, v \notin N_G(w)\}$, and $t_{uv} = |T_{uv}|$. If u and v are clearly understood, we sometimes write λ instead of λ_{uv}, T instead of T_{uv}, and t

instead of t_{uv}. The cardinality of a maximum independent set containing u and v is denoted by $\alpha_{uv}(G)$.

THEOREM 19. (Ainouche and Christofides [1]). *Let u, v be two nonadjacent vertices of a graph G such that*

$$\alpha_{uv} \leq \lambda_{uv}; \tag{2}$$

then G is Hamiltonian if and only if $G + uv$ is Hamiltonian.

THEOREM 20 (Ainouche and Christofides [2]). *Let u, v be two nonadjacent vertices of a 2-connected graph G such that*

$$d_G(w) \geq t + 2 \quad \textit{for at least } \min\{t, t + 2 - \lambda\} \textit{ vertices } w \in T; \tag{3}$$

then G is Hamiltonian if and only if $G + uv$ is Hamiltonian.

Note that both Theorems 19 and 20 generalize Theorem 18.

It was observed in Broersma and Schiermeyer [11] that (3) can be restated in terms of degree sums of independent triples.

PROPOSITION 3. *Equation* (3) *is equivalent to*

$$d(u) + d(v) + d(w) \geq n + \lambda_{uv} \quad \textit{for at least } \min\{t, t + 2 - \lambda_{uv}\} \textit{ vertices } w \in T. \tag{4}$$

Our next observation admits a corollary useful in the proof of Theorem 4.

PROPOSITION 4. *Equation* (4) *is equivalent to*

$$|N(u) \cup N(v)| \geq n - d(w) \quad \textit{for at least } \min\{t, t + 2 - \lambda_{uv}\} \textit{ vertices } w \in T. \tag{5}$$

Proof. This follows directly from $d(u) + d(v) = |N(u) \cup N(v)| + \lambda_{uv}$. □

COROLLARY 2. *Let u, v be two nonadjacent vertices of a 2-connected graph G such that $|N(u) \cup N(v)| \geq n - \delta$; then G is Hamiltonian if and only if $G + uv$ is Hamiltonian.*

The next theorem is due to Bigalke and Jung [6].

THEOREM 21. *Let G be a 1-tough graph with $n \geq 3$ vertices. If*

$$\delta \geq \max\left\{\frac{n}{3}, \alpha - 1\right\},$$

then G is Hamiltonian.

We need a similar statement involving σ_2 instead of δ. This can be obtained by using the inequality $\frac{1}{3}\sigma_3 \geq \frac{1}{2}\sigma_2$ from the next theorem.

THEOREM 22 (Bauer et al. [5]). *Let G be a 1-tough graph with $n \geq 3$ vertices. If $\sigma_3 \geq n$, then G contains a cycle of length at least $\min\{n, n + \frac{1}{3}\sigma_3 - \alpha\}$.*

COROLLARY 3. *Let G be a 1-tough graph with $n \geq 3$ vertices. If*

$$\tfrac{1}{2}\sigma_2 \geq \max\left\{\tfrac{1}{3}n, \alpha - \tfrac{1}{2}\right\},$$

then G is Hamiltonian.

For disjoint subsets $A, B \subseteq V(G)$, let $\mathrm{odd}(A, B)$ denote the number of components H of the graph $G - (A \cup B)$ with $e(H, B)$ odd and let $\Theta(A, B) = 2|A| + \sum_{v \in B} d_{G-A}(v) - 2|B| - \mathrm{odd}(A, B)$. The following theorem of Tutte [25] characterizes those graphs that do not contain a 2-factor.

THEOREM 23. *Let G be a graph. Then* (i) $\Theta(A, B)$ *is even for all disjoint subsets* A, B *of* $V(G)$, *and* (ii) G *does not have a* 2*-factor if and only if there exist disjoint subsets* A, B *of* $V(G)$ *with* $\Theta(A, B) \leq -2$.

4. Proofs.

Proof of Theorem 6. Let G be a non-Hamiltonian graph with $\delta + \Delta \geq n$. We must show that G is not path-tough. Let $u \in V(G)$ be a vertex with $d(u) = \Delta$. Then $d(u) + d(v) \geq \Delta + \delta \geq n$ holds for every vertex $v \in V(G)$. Therefore, by repeated application of Theorem 18, we have that G is Hamiltonian if and only if $H = G + \{uv | v \notin N_G(u) \cup \{u\}\}$ is Hamiltonian. Hence H is not Hamiltonian. Clearly, $d_H(u) = n - 1$, and thus $H - u = G - u$ has no Hamilton path; that is, G is not path-tough. □

Proof of Theorem 4. Suppose that there is a non-Hamiltonian path-tough graph G on n vertices with $\delta \geq [3/(6 + \sqrt{3})]n$. We may assume that G is maximal non-Hamiltonian. Thus, each pair of nonadjacent vertices does not satisfy any of the conditions of Theorems 18–20 or Corollary 2, respectively. Theorem 6 implies that

$$\Delta \leq n - \delta - 1. \tag{6}$$

Since G is path-tough, we have $\delta \geq 2$ and $\alpha \leq \frac{1}{2}n$. With Theorem 21, we obtain

$$4 \leq \delta + 2 \leq \alpha \leq \frac{n}{2}. \tag{7}$$

Let I be an independent set of size α and let $U = V(G) - I$. If $G[I, U]$ is complete, then $d(v) = n - \alpha \geq n/2$ for all vertices $v \in I$, contradicting Theorem 18. Hence, we may assume that

$$G\,[I, U] \text{ is not complete.} \tag{8}$$

Claim 1. If $u \in U$ and $v \in I$ are not adjacent, then $|N_I(u)| \leq n - 2\delta - 1$.

By Corollary 2, we have $|N(u) \cup N(v)| \leq n - \delta - 1$, and so $|N_I(u)| \leq n - 2\delta - 1$, since $N_I(v) = \varnothing$ implies that $|N_U(v)| \geq \delta$.

Now let $U_1 = \{u \in U | |N_I(u)| \geq n - 2\delta\}$ and $U_2 = U - U_1$. Then, by Claim 1,

$$G\,[I, U_1] \text{ is complete.} \tag{9}$$

Furthermore, $U_2 \neq \varnothing$ by (8) and Claim 1. Let $m = |U_2|$. Then $1 \leq m \leq n - \alpha$.

Claim 2. $U_1 \neq \varnothing$ and $G[U_1]$ is complete.

Suppose first that $U_1 = \varnothing$. Then we have with (7)

$$\delta\,(\delta + 2) \leq \delta\alpha \leq e\,(U, I) \leq (n - 2\delta - 1)\,(n - \alpha) \leq (n - 2\delta - 1)\,(n - \delta - 2)\,,$$

implying that $0 \leq n^2 - (3\delta + 3)n + (\delta + 1)(\delta + 2)$. This, however, is impossible for $[3/(6 + \sqrt{3})]n \leq \delta \leq \frac{1}{2}n$. Suppose now that $x, y \in U_1$ are nonadjacent. Then $\alpha_{xy} \leq \alpha \leq \lambda_{xy}$ by (9), contradicting Theorem 19.

Now let $U_{21} = \{u \in U_2 | |N_{U_2}(u)| \geq m - \delta + 1\}, U_{22} = U_2 - U_{21}$, and $p = |U_{21}|$.

Claim 3. $G[U_1, U_{21}]$ is complete.

Suppose that $x \in U_1$ and $y \in U_{21}$ are nonadjacent. Then $|N(x) \cup N(y)| \geq |N_{U_1 \cup I}(x)| + |N_{U_2}(y)| \geq ((n - \alpha - m - 1) + \alpha) + (m - \delta + 1) = n - \delta$, contradicting Corollary 2.

Claim 4. $m - p \geq \delta$.

If $m - p \leq \delta - 1$, then we have $d(u) \geq (n - \alpha - m - 1) + \alpha + p = n - (m - p) - 1 \geq n - \delta$ for all $u \in U_1$, contradicting (6) and Claim 2.

Claim 5. There exists a vertex $u \in U_1$ satisfying

$$d(u) \geq n - (m - p) - 1 + \frac{(m-p)(4\delta - n - m + 1)}{n - m - \alpha}.$$

For all $v \in U_{22}$, we have

$$|N_{U_1}(v)| \geq \delta - |N_I(v)| - |N_{U_2}(v)| \geq \delta - (n - 2\delta - 1) - (m - \delta) = 4\delta - n + 1 - m,$$

and so there exists $u \in U_1$ with

$$|N_{U_{22}}(u)| \geq \frac{(m-p)(4\delta - n + 1 - m)}{n - m - \alpha}.$$

Therefore, by (9) and Claims 2 and 3,

$$\begin{aligned} d(u) &= |N_I(u)| + |N_{U_1}(u)| + |N_{U_{21}}(u)| + |N_{U_{22}}(u)| \\ &\geq \alpha + (n - m - \alpha - 1) + p + \frac{(m-p)(4\delta - n - m + 1)}{n - m - \alpha}. \end{aligned}$$

Claim 6. $m < [(\sqrt{3} + 1)/2]\delta - 2$.

Let $l = |U_1|$. We have

$$\begin{aligned} \alpha\delta &\leq e(I, U) = e(I, U_1) + e(I, U_2) \\ &\leq \alpha l + m(n - 2\delta - 1) = \alpha l + (n - \alpha - l)(n - 2\delta - 1). \end{aligned}$$

Equivalently, since $\alpha + 2\delta - n + 1 \geq 3\delta - n + 3 > 0$ by Theorem 21,

$$l \geq \frac{\alpha\delta - (n - \alpha)(n - 2\delta - 1)}{\alpha + 2\delta - n + 1}.$$

Denote the right-hand side of the last inequality by $f_1(\alpha)$. Then $f_1'(\alpha) = (n - 2\delta - 1)(\delta + 1)/(\alpha + 2\delta - n + 1)^2 > 0$ by (7). Therefore, f_1 is monotone increasing in α, and so again by (7)

$$l \geq f_1(\delta + 2) = \frac{n\delta - (n - \delta - 2)(n - \delta - 1)}{3\delta - n + 3} > \frac{n\delta - (n - \delta)^2}{3\delta - n},$$

where the last estimate follows from $n > 2\delta + 2\delta/(3\delta + 2)$, which is an immediate consequence of (7). Now let $f_2(n) = (n\delta - (n - \delta)^2)/(3\delta - n)$. Then $f_2'(n) = ((n - 3\delta)^2 - \delta^2)/(3\delta - n)^2 < 0$ holds, since $2\delta < n < 3\delta$. Therefore, f_2 is strictly decreasing in n, and

$$f_2(n) \geq f_2\left(\frac{6 + \sqrt{3}}{3}\delta\right) = \frac{2 - \sqrt{3}}{3 - \sqrt{3}}\delta.$$

Together with (7), we obtain

$$m = n - \alpha - l < \frac{6 + \sqrt{3}}{3}\delta - (\delta + 2) - \frac{2 - \sqrt{3}}{3 - \sqrt{3}}\delta = \frac{\sqrt{3} + 1}{2}\delta - 2.$$

Therefore, Claim 6 is proved.

Now we complete the proof of the theorem as follows. By Claim 5 and (6), we have

$$\begin{aligned} n-\delta-1 &\geq n-(m-p)-1+\frac{(m-p)(4\delta-n-m+1)}{n-m-\alpha} \\ &\geq n-(m-p)-1+\frac{(m-p)(4\delta-n-m+1)}{n-m-\delta-2}, \end{aligned} \tag{10}$$

where the last estimate follows from (7) and from $4\delta-n-m+1>4\delta-[(6+\sqrt{3})/3]\delta-[(\sqrt{3}+1)/2]\delta=[(9-5\sqrt{3})/6]\delta>0$, which holds by Claim 6. Now (10) is equivalent to

$$\delta \leq \frac{(m-p)(2n-5\delta-3)}{n-m-\delta-2},$$

which implies, together with $m-p>0$, that $(2n-5\delta-3)/(n-m-\delta-2)>0$, and thus

$$\delta \leq m\frac{(2n-5\delta-3)}{n-m-\delta-2}.$$

Let $f_3(n)=(2n-5\delta-3)/(n-m-\delta-2)$. Then $f_3'(n)=(3\delta-2m-1)/(n-m-\delta-2)^2>0$ by Claim 6. So f_3 is monotone increasing in n; hence

$$\delta \leq mf_3\left(\frac{6+\sqrt{3}}{3}\delta\right)=m\frac{(2\sqrt{3}-3)\delta/3-3}{(3+\sqrt{3})\delta/3-m-2}.$$

Since we have

$$m+2<\frac{\sqrt{3}+1}{2}\delta<\frac{\sqrt{3}+3}{3}\delta$$

by Claim 6, the denominator and numerator of the fraction on the right-hand side of the last equality are positive. Finally, using Claim 6, we obtain

$$\delta<\left(\frac{\sqrt{3}+1}{2}\delta-2\right)\frac{(2\sqrt{3}-3)\delta/3-3}{(3+\sqrt{3})\delta/3-(\sqrt{3}+1)\delta/2}<\frac{\sqrt{3}+1}{2}\delta\frac{(2\sqrt{3}-3)\delta/3}{(3-\sqrt{3})\delta/6}=\delta,$$

a contradiction. □

Proof of Theorem 8. Let G be a 2-connected, non-Hamiltonian graph satisfying $\frac{1}{2}\sigma_2+\Delta\geq n$. We must show that G is not path-tough.

Let $u\in V(G)$ be a vertex with $d(u)=\Delta$ and let v be a further vertex not adjacent to u. If $T_{uv}=\varnothing$, then (3) of Theorem 20 is satisfied. If $d(v)\geq\frac{1}{2}\sigma_2$, then $d(u)+d(v)\geq\Delta+\frac{1}{2}\sigma_2\geq n$, and therefore (1) of Theorem 18 is satisfied. Finally, if $T_{uv}\neq\varnothing$ and $d(v)<\frac{1}{2}\sigma_2$, then $d(w)\geq\sigma_2-d(v)>\frac{1}{2}\sigma_2$ for all vertices $w\in T_{uv}$. Therefore (1) of Theorem 18 is satisfied for u and every vertex $w\in T_{uv}$. Thus it follows by repeated application of Theorems 18 and 20 that G is Hamiltonian if and only if $H=G+\{ux|x\notin N_G(u)\cup\{u\}\}$ is Hamiltonian. Now the proof can be completed in the same way as the proof of Theorem 6. □

Proof of Theorem 7. The proof follows the lines of the proof of Theorem 4. Again, we suppose that G is a non-Hamiltonian path-tough graph on n vertices with $\sigma_2\geq\frac{4}{5}(n-1)$ and that G is maximal non-Hamiltonian. The notation is the same as above, except with δ replaced by $\sigma_2/2$. I is an independent set of size $\alpha, U=V(G)-I, U_1=$

$\{u \in U \| N_I(u)| \geq n-\sigma_2\}, U_2 = U - U_1$, and $m = |U_2|$. The following statements (11)–(13) are proved analogously to the corresponding statements in the proof of Theorem 4. Use Theorem 8 to obtain (11) and Corollary 3 for (12). We have that

$$\Delta < n - \frac{\sigma_2}{2}, \tag{11}$$

$$3 \leq \frac{1}{2}\sigma_2 + 1 \leq \alpha \leq \frac{n}{2}, \tag{12}$$

$$G\,[I, U] \text{ is not complete.} \tag{13}$$

Next, we show that

$$G\,[I, U_1] \text{ is complete.} \tag{14}$$

Therefore we suppose that $u \in U_1$ and $v \in I$ are not adjacent. By Proposition 4, we have $|N(u) \cup N(v)| < n - d(w)$ for at least one vertex $w \in T_{uv}$. Thus we obtain $|N_I(u)| \leq |N(u) \cup N(v)| - d(v) < n - (d(w) + d(v)) \leq n - \sigma_2$, contradicting $u \in U_1$.

As in Claim 2 in the proof of Theorem 4, we obtain the facts that

$$U_1 \neq \varnothing \quad \text{and} \quad G\,[U_1] \text{ is complete.} \tag{15}$$

We note that

$$m \geq \frac{\sigma_2 - 1}{2}, \tag{16}$$

since otherwise $|U_1| \geq n - \alpha - \frac{1}{2}\sigma_2 + 1$ and any vertex $w \in U_1$ satisfies

$$\Delta \geq d_{U_1}(w) + d_I(w) = |U_1| - 1 + \alpha \geq n - \alpha - \frac{\sigma_2}{2} + \alpha = n - \frac{\sigma_2}{2},$$

contradicting (11).

From $\alpha \geq 2$, we have $\frac{1}{2}\alpha\sigma_2 \leq e(I, U)$, and thus

$$\alpha\frac{\sigma_2}{2} \leq e\,(I, U_1) + e\,(I, U_2) \leq \alpha\,(n - \alpha - m) + m\,(n - \sigma_2 - 1)\,.$$

By (12), the right-hand side of the above inequality is decreasing in m, since the hypothesis of the theorem implies that $-\alpha + n - \sigma_2 - 1 \leq n - \frac{3}{2}\sigma_2 - 2 \leq n - \frac{3}{2}(\frac{4}{5}(n - 1)) - 2 < 0$ holds. Hence by (16)

$$0 \leq \alpha\left(n - \alpha - \frac{\sigma_2 - 1}{2}\right) + \frac{\sigma_2 - 1}{2}\,(n - \sigma_2 - 1) - \alpha\frac{\sigma_2}{2}.$$

The last expression is decreasing in α, since its derivative with respect to α is $n - 2\alpha - \sigma_2 + \frac{1}{2} \leq n - 2\sigma_2 - \frac{3}{2} < 0$ by (12) and the hypothesis of the theorem. Thus (12) yields

$$\begin{aligned} 0 &\leq \left(\frac{\sigma_2}{2} + 1\right)\left(n - \sigma_2 - \frac{1}{2}\right) + \frac{\sigma_2 - 1}{2}\,(n - \sigma_2 - 1) - \left(\frac{\sigma_2}{2} + 1\right)\frac{\sigma_2}{2} \\ &= -\frac{5}{4}\sigma_2^2 + \left(n - \frac{7}{4}\right)\sigma_2 + \frac{n}{2} \end{aligned}$$

or, equivalently,

$$\tfrac{2}{5}\left(n-\tfrac{7}{4}\right)-\tfrac{1}{5}\sqrt{4\left(n-\tfrac{7}{4}\right)^2+10n}\le\sigma_2\le\tfrac{2}{5}\left(n-\tfrac{7}{4}\right)+\tfrac{1}{5}\sqrt{4\left(n-\tfrac{7}{4}\right)^2+10n}.$$

This already contradicts $\sigma_2 \geq \frac{4}{5}(n-1)$, if $n > 6$, and yields equality in all above estimates for $n = 6$. In the latter case, we especially have $\sigma_2 = 4$, implying that $m = \frac{3}{2}$ by (16), also a contradiction. Therefore we have $n \leq 5$, contradicting (12). □

Proof of Theorem 14. Suppose that G satisfies the hypothesis of Theorem 14 but contains no 2-factor. Suppose furthermore that G is chosen edge-maximal with these properties. By Theorem 23, there exist disjoint subsets A, B of $V(G)$ with $\Theta(A,B) \leq -2$. We choose A and B such that $\Theta(A,B') \geq 0$ for each proper subset $B' \subset B$.

Claim 7. $A = \varnothing$.

Suppose that A contains a vertex x. Since addition of any missing edge incident with x leaves $\Theta(A,B)$ unchanged, the edge-maximality of G implies that x is joined to all other vertices of G. Now it follows from the path-toughness of G that $G-x$ contains a Hamilton path, and therefore G is Hamiltonian, a contradiction.

Claim 8. B is an independent set.

(Note that this follows from a result in [15]; we include the proof for completeness.)

We show that $e(x, B-x) = 0$ for any $x \in B$. By the choice of B, we have $\Theta(\varnothing, B-x) \geq 0$, and thus $-2 \geq \Theta(\varnothing,B) - \Theta(\varnothing, B-x) = d_G(x) + \mathrm{odd}(\varnothing, B-x) - \mathrm{odd}(\varnothing,B) - 2$, which implies that

$$\mathrm{odd}\,(\varnothing,B)-\mathrm{odd}\,(\varnothing,B-x)\ge d_G\,(x)\,. \tag{17}$$

Moreover, $\mathrm{odd}(\varnothing, B-x) \geq \mathrm{odd}(\varnothing,B) - e(x, V(G)-B)$ holds, or, equivalently,

$$\mathrm{odd}\,(\varnothing,B-x)-\mathrm{odd}\,(\varnothing,B)\ge -e\,(x,V\,(G)-B)\,. \tag{18}$$

Adding (17) and (18) gives

$$0\ge d_G\,(x)-e\,(x,V\,(G)-B)=e\,(x,B-x)\ge 0.$$

This proves the claim.

Now we consider two cases.

Case 1. $|B| \geq 3$.

$\Theta(\varnothing,B) \leq -2$ is equivalent to

$$\mathrm{odd}\,(\varnothing,B)\ge 2+\sum_{v\in B}d_G\,(v)-2\,|B|\,,$$

and $\sigma_3(G) \geq 7$ yields $\sum_{v\in B} d_G(v) \geq 3|B|-2$ by Claim 8 and $|B| \geq 3$. Therefore we have $\mathrm{odd}(\varnothing,B) \geq |B|$. On the other hand, the path-toughness of G yields $|B| \geq \omega(G-B) \geq \mathrm{odd}(\varnothing,B)$, implying that $|B| = \omega(G-B) = \mathrm{odd}(\varnothing,B)$ and $\sum_{v\in B} d_G(v) = 3|B|-2$.

Now let $H_1, H_2, \ldots, H_{|B|}$ denote the components of $G-B$. Then we have $e(H_i,B) \geq 3$ for any $i = 1,\ldots,|B|$, since G is 2-connected and $e(H_i,B)$ is odd. So, with Claim 8, we obtain

$$3\,|B|-2=\sum_{v\in B}d_G\,(v)=\sum_{i=1}^{|B|}e(H_i,B)\ge 3\,|B|\,.$$

This contradiction completes the discussion of this case.

CASE 2. $|B| \leq 2$.

Since $\Theta(\varnothing, \varnothing) = 0$, we see that $|B| = 1$ or $|B| = 2$. If $|B| = 1$, say $B = \{x\}$, then $\Theta(\varnothing, \{x\}) \leq -2$ and $d_G(x) \geq 2$ imply that $\text{odd}(\varnothing, \{x\}) \geq 2$. Therefore G is not 2-connected, a contradiction.

Finally, if $|B| = 2$, say $B = \{x_1, x_2\}$, then we obtain from $\Theta(\varnothing, B) \leq -2$ that

$$\text{odd}\,(\varnothing, B) \geq 2 + d_G\,(x_1) + d_G\,(x_2) - 4 \geq 2.$$

Also, since G is 1-tough, we have

$$2 \geq \omega\,(G - B) \geq \text{odd}\,(\varnothing, B)\,.$$

Thus $\omega(G - B) = \text{odd}(\varnothing, B) = 2$ and $d_G(x_1) = d_G(x_2) = 2$. This is impossible, however, since, for both components H_1, H_2 of $G - B$, we have $e(H_i, B) \geq 3, i = 1, 2$, as in Case 7. This contradiction completes the proof of the theorem. □

Proof of Theorem 15. We transform HAMILTON PATH to PATH-TOUGH. Let the graph G be an arbitrary instance of HAMILTON PATH. We now consider the graph $G' = K_1 + G$. Clearly, it is possible to construct G' from G in polynomial time.

We now show that G has a Hamilton path if and only if G' is path-tough. Clearly, if G' is path-tough, then G has a Hamilton path. Suppose now that G has a Hamilton path, say $v_1 v_2 \cdots v_n$, where n is the order of G. Let u be the unique vertex in $V(G') - V(G)$. Then $G = G' - u$ has the Hamilton path $v_1 v_2 \cdots v_n$, and, for every $1 \leq i \leq n, G' - v_i$ has the Hamilton path $v_1 v_2 \cdots v_{i-1} u v_{i+1} \cdots v_n$. Therefore G' is path-tough by Proposition 2. □

Proof of Theorem 16. We transform HAMILTON PATH to HAMILTON CYCLE ($\delta \geq (\frac{1}{2} - \varepsilon)n$). Let $G_1 = (V_1, E_1)$ be a graph with n_1 vertices making up an arbitrary instance of HAMILTON PATH. We now construct a graph $G_2 = (V_2, E_2)$ by adding $2p - 1$ vertices $v_1, v_2, \ldots, v_{2p-1}$, where $p = \lceil (\frac{1}{2} - \varepsilon) n_2 \rceil$ and $n_2 = \lceil n_1/(2\varepsilon) \rceil$. Then $V_2 = V_1 \cup \{v_1, v_2, \ldots, v_{2p-1}\}$. Furthermore, let $E_2 = E_1 \cup \{v_i v_j | 1 \leq i \leq p, p+1 \leq j \leq 2p - 1\} \cup \{v_i v | v \in V_1, 1 \leq i \leq p\}$. Thus $\delta(G_2) = p \geq (\frac{1}{2} - \varepsilon) n_2$. For each fixed $\varepsilon > 0$, the graph G_2 has size $O(n_2^2)$, which is bounded above by a polynomial function of n_1. Therefore, it is possible to construct G_2 from G_1 in polynomial time.

If G_2 has a Hamilton cycle, then G_1 has a Hamilton path, since $G_2 - \{v_1, v_2, \ldots, v_p\}$ has p components $v_{p+1}, \ldots, v_{2p-1}$ and G_1. If G_1 has a Hamilton path, say $u_1 u_2 \cdots u_{n_1}$, then $u_1 u_2 \cdots u_{n_1} v_1 v_{p+1} v_2 v_{p+2} \cdots v_{2p-1} v_p u_1$ is a Hamilton cycle in G_2. Thus G_1 has a Hamilton path if and only if G_2 has a Hamilton cycle. □

5. Concluding remarks. The circumference $c(G)$ of a graph G is the length of a longest cycle. Clearly, a graph G of order n is Hamiltonian if and only if $c(G) = n$.

In view of Conjecture 1, we obtained lower bounds for the circumference of path-tough graphs satisfying a σ_3-condition. Our proofs distinguish two situations. First, if the graph is traceable, then, instead of "path-tough," only "2-connected" is needed to prove the following result.

THEOREM 24. *Let G be a 2-connected graph with n vertices. If G is traceable and $\sigma_3 \geq n + 2$, then $c(G) \geq n - 1$.*

Second, if the graph is nontraceable, we obtain Theorem 25.

THEOREM 25. *Let G be a path-tough graph with n vertices. If G is nontraceable and $\sigma_3 \geq n$, then every vertex of G is contained in a cycle of length $c(G) = n - 2$.*

Note that path-tough, nontraceable graphs are usually called *hypotraceable.* We are aware of the problem that possibly no graph satisfies the hypotheses of Theorem 25, since no hypotraceable graphs with $\delta \geq 4$ are known (cf. [10]).

These theorems have the following corollary.

COROLLARY 4. *Let G be a path-tough graph with n vertices. If $\sigma_3 \geq n+2$, then $c(G) \geq n-2$.*

Hendry [21] posed the problem of determining the maximum number $f(n)$ of edges in a non-Hamiltonian path-tough graph of order n. He conjectured that there exist integers N and c such that

$$f(n) \leq \binom{n-7}{2} + c \quad \text{for } n \geq N.$$

This conjecture is disproved by the graph G_{pqr} with $p = n-6$ and $q = r = 3$. Therefore we see that $f(n) \geq \binom{n-6}{2} + 12$.

Chvátal [12] introduced toughness as a parameter of a graph, and, of course, it is also possible to introduce the *path-toughness* $\pi(G)$ of a graph G. We suggest the following:

$$\pi(G) = \min\left\{ \frac{|S|}{\mu(G-S)} | S \neq \varnothing, \mu(G-S) \geq 2 \right\}.$$

(Note that here $S = \varnothing$ is excluded to avoid the existence of path-tough graphs with path-toughness 0.) Clearly, the following holds.

PROPOSITION 5. *Let G be a connected graph. Then $\pi(G) \leq t(G)$.*

One motivation for the study of $\pi(G)$ is Chvátal's well-known conjecture that there exists a constant t_0 such that every graph G with $t(G) \geq t_0$ is Hamiltonian, because this conjecture implies the next one.

CONJECTURE 2. *There exists a constant t_1 such that every graph G with $\pi(G) \geq t_1$ is Hamiltonian.*

Furthermore, it would be interesting to know how large the difference $t(G) - \pi(G)$ can be. For example, if someone can prove that there exists a constant c such that $t(G) - \pi(G) \leq c$ for all graphs G, then it turns out that Chvátal's conjecture and Conjecture 2 are equivalent. The following examples show that c must be at least $\frac{3}{2}$. The *corona* $G \circ K_1$ of a graph G is the graph obtained from G by joining every vertex of G to exactly one new vertex. For positive integers a, b, c, let $G(a,b,c) = K_a + b(K_c \circ K_1)$. We may verify that

$$t(G(a,b,c)) = \begin{cases} \frac{a}{b} & \text{if } a < b, \\ \frac{a+(c-1)b}{bc} & \text{if } a \geq b, \end{cases}$$

and, for c odd,

$$\pi(G(a,b,c)) = \begin{cases} (b(c+1)/2 - a + 1)^{-1} & \text{if } a \leq b(c+1)/2 - 1, \\ \frac{2a}{b(c+1)} & \text{if } b(c+1)/2 \leq a \leq (c+1)b, \\ \frac{a+(c-1)b}{bc} & \text{if } a > (c+1)b. \end{cases}$$

Therefore, if we choose integers $d \geq 1, b \geq 1$, and $c \geq 1$, odd, with $b(c-1) > 2d$, and let $a = b(c+1)/2 - d$, we have $\pi(G(a,b,c)) = 1/(d+1)$ and $t(G(a,b,c)) = \frac{3}{2} - (b+2d)/2bc$. Thus, choosing d and c large, the difference $t - \pi$ can be made arbitrarily close to $\frac{3}{2}$. If we restrict our attention only to path-tough graphs and let $a = b(c+1)/2$, we obtain $\pi(G(a,b,c)) = 1$ and $t(G(a,b,c)) = \frac{3}{2} - 1/2c$. So, for path-tough graphs, the difference

$t - \pi$ can be arbitrarily close to $\frac{1}{2}$. Moreover, note that, for the graphs $G(a, b, c)$, the toughness and the path-toughness are equal if the toughness is at least 2. In fact, we do not know any graph G with $t(G) \geq 2$ and $t(G) > \pi(G)$.

Finally, we think it interesting to investigate bipartite graphs, and so we ask if there exist bipartite path-tough graphs that are not Hamiltonian.

Notes added in proof. Theorem 16 is also proved in Häggkvist [20]. Recently, J. van den Heuvel informed us about the following results in Enomoto et al. [14]. (See also van den Heuvel [22].)

THEOREM 26. *Let G be a connected graph on $n \geq 3$ vertices such that $\sigma_3 \geq n$. Then, for every path P in G, there exists a cycle C in G such that $|V(P) - V(C)| \leq 1$ or G belongs to a set $\mathcal{F}(n)$ of non-1-tough graphs.*

This theorem implies Theorem 24 even under the weaker condition $\sigma_3 \geq n$. Therefore, Corollary 4 remains true if $\sigma_3 \geq n + 2$ is replaced by $\sigma_3 \geq n$.

Finally, H. Müller from Jena (personal communication) informed us about the existence of a bipartite, path-tough, non-Hamiltonian graph. His example consists of two vertex disjoint cycles of length 6, say $v_1, v_2, v_3, v_4, v_5, v_6$ and $u_1, u_2, u_3, u_4, u_5, u_6$ and the additional edges v_1u_1, v_2u_4, v_4u_2, and v_5u_5.

REFERENCES

[1] A. AINOUCHE AND N. CHRISTOFIDES, *Strong sufficient conditions for the existence of Hamiltonian cycles in undirected graphs*, J. Combin. Theory Ser. B, 31 (1981), pp. 339–343.

[2] ———, *Semi-independence number of a graph and the existence of Hamiltonian circuits*, Discrete Appl. Math., 17 (1987), pp. 213–221.

[3] D. BAUER, H. J. BROERSMA, L. RAO, AND H. J. VELDMAN, *A generalization of a result of Häggkvist and Nicoghossian*, J. Combin. Theory Ser. B, 47 (1989), pp. 237–243.

[4] D. BAUER, S. L. HAKIMI, AND E. SCHMEICHEL, *Recognizing tough graphs is* NP-*hard*, Discrete Appl. Math., 28 (1990), pp. 191–195.

[5] D. BAUER, E. F. SCHMEICHEL, A. MORGANA, AND H. J. VELDMAN, *Long cycles in graphs with large degree sums*, Discrete Math., 79 (1989/90), pp. 59–70.

[6] A. BIGALKE AND H. A. JUNG, *Über Hamiltonsche Kreise und unabhängige Ecken in Graphen*, Monatsh. Math., 88 (1979), pp. 195–210.

[7] F. T. BOESCH, S. CHEN, AND J. A. M. MCHUGH, *On covering the points of a graph with point disjoint paths*, in Graphs and Combinatorics, Lecture Notes in Math., No. 406, 1974, pp. 201–212.

[8] J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, Macmillan, London, 1976.

[9] J. A. BONDY, *Longest Paths and Cycles in Graphs of High Degree*, Research Report CORR 80-16, University of Waterloo, Waterloo, Ontario, Canada, 1980.

[10] ———, *Basic graph theory: Paths and circuits*, in Handbook of Combinatorics, R. L. Graham, M. Grötschel, and L. Lovász, eds., North–Holland, to appear.

[11] H. J. BROERSMA AND I. SCHIERMEYER, *A closure concept based on neighborhood unions of independent triples*, Discrete Math., 124 (1993), pp. 37–47.

[12] V. CHVÁTAL, *Tough graphs and Hamiltonian circuits*, Discrete Math., 5 (1973), pp. 215–228.

[13] G. A. DIRAC, *Some theorems on abstract graphs*, Proc. London Math. Soc., 2 (1952), pp. 69–81.

[14] H. ENOMOTO, J. VAN DEN HEUVEL, A. KANEKO, AND A. SAITO, *Relative Length of Long Paths and Cycles in Graphs with Large Degree Sums*, preprint, 1993.

[15] H. ENOMOTO, B. JACKSON, P. KATERINIS, AND A. SAITO, *Toughness and the existence of k-factors*, J. Graph Theory, 9 (1985), pp. 87–95.

[16] B. FAßBENDER, *A Sufficient Condition on Degree Sums of Independent Triples for Hamiltonian Cycles in* 1-*Tough Graphs*, Ars Combin., 33 (1992), pp. 300–304.

[17] M. GAREY AND D. JOHNSON, *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[18] S. Goodman and S. Hedetniemi, *On the Hamiltonian completion problem*, in Graphs and Combinatorics, Lecture Notes in Math., No. 406, 1974, pp. 262–272.

[19] R. Häggkvist, *Twenty odd statements in Twente*, presented at the Twente Workshop on Hamiltonian Graph Theory, 1992; announced in Ann. Discrete Math., 9 (1980), p. 259.

[20] R. Häggkvist, *On the structure of non-Hamiltonian graphs* I. Combin. Probab. Comput., 1 (1992), pp. 27–34.

[21] G. R. T. Hendry, *Scattering number and extremal non-Hamiltonian graphs*, Discrete Math., 71 (1988), pp. 165–175.

[22] J. van den Heuvel, *Degree and Toughness Conditions for Cycles in Graphs*, Ph.D. dissertation, Univ. of Twente, Enschede the Netherlands, 1993.

[23] H. A. Jung, *On maximal circuits in finite graphs*, Ann. Discrete Math., 3 (1978), pp. 129–144.

[24] O. Ore, *Note on Hamiltonian circuits*, Amer. Math. Monthly, 67 (1960), p. 55.

[25] W. T. Tutte, *The factors of graphs*, Canad. J. Math., 4 (1952), pp. 314–328.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 585–592, November 1994

A NEW CLASS OF PYRAMIDALLY SOLVABLE SYMMETRIC TRAVELING SALESMAN PROBLEMS*

JACK A. A. VAN DER VEEN[†]

Abstract. An instance of the symmetric traveling salesman problem (STSP) is pyramidally solvable if there is a shortest tour that is pyramidal. A pyramidal tour is a Hamiltonian tour that consists of two parts; according to the labeling of the vertices in the first part the vertices are visited in increasing order and in the second part in decreasing order. It is well known that a shortest pyramidal tour can be found in $\mathcal{O}(n^2)$ time. In this paper it is shown that the STSP restricted to the class of distance matrices

$$\mathbb{D}_{\text{NEW}} = \left\{ D = (d[i,j]) \,\middle|\, \begin{array}{ll} d[i,j] = d[j,i] & \text{for all } i \text{ and } j \\ d[i,j] + d[j+1,k] \leq d[i,k] + d[j,j+1] & \text{for all } i < j < j+1 < k \end{array} \right\}$$

is pyramidally solvable. Furthermore, it is shown that $\mathbb{D}_{\text{NEW}} \not\subset \mathbb{D}_{\text{DEMI}}$, i.e., that $\mathbb{D}_{\text{NEW}}$ is not contained in the class of symmetric Demidenko matrices $\mathbb{D}_{\text{DEMI}}$ that was, until now, the most general class of pyramidally solvable STSPs. It is also shown that $\mathbb{D}_{\text{DEMI}} \not\subset \mathbb{D}_{\text{NEW}}$.

Key words. traveling salesman problem, solvable case, pyramidal tour

AMS subject classification. 90C99

1. Introduction.

The *traveling salesman problem* (TSP) is the problem of finding a shortest Hamiltonian cycle (tour) in a complete weighted digraph $G = (N, A, d)$ with set of vertices $N = \{1, \ldots, n\}$ (throughout we will assume $n \geq 4$), set of arcs $A = \{[i,j] | i, j \in N\}$, and $D = (d[i,j])$ the matrix with distances associated with the arcs in A. A shortest Hamiltonian cycle is a cyclic permutation φ of N that solves

$$[\text{TSP}] \quad \text{Minimize} \left\{ d(\varphi) := \sum_{i=1}^{n} d[i, \varphi(i)] \,\middle|\, \varphi \text{ is a Hamiltonian tour} \right\}.$$

Obviously, $\varphi(i)$ denotes the successor of a vertex i in the cyclic permutation φ. A more general notation is useful: for any natural number z, we denote the zth successor (zth predecessor) of i by $\varphi^z(i)(\varphi^{-z}(i)$, respectively), so that $\varphi^1(i) = \varphi(i)$ and, with $\varphi^0(i) = i, \varphi^{s+t}(i) = \varphi^s(\varphi^t(i))$ for all integers s and t. A Hamiltonian cycle φ will be denoted by

$$(\varphi^0(1), \varphi^1(1), \ldots, \varphi^{n-1}(1))$$

so that $\varphi = (i_1, i_2, \ldots, i_n)$ is to be interpreted as $i_1 = 1, \varphi(i_k) = i_{k+1}, k = 1, \ldots, n-1$, and $\varphi(i_n) = 1$.

A *pyramidal tour* is a Hamiltonian tour of the form

$$(1, i_1, i_2, \ldots, i_r, n, j_1, j_2, \ldots, j_{n-r-2})$$

with $i_1 < i_2 < \cdots < i_r$ and $j_1 > j_2 > \cdots > j_{n-r-2}$. A useful characterization of pyramidal tours is the following. A vertex v is called a *peak* of a (cyclic) permutation

* Received by the editors December 3, 1990; accepted for publication (in revised form) October 8, 1993.

[†] Department of Quantitative Methods, Nijenrode University, Straatweg 25, 3621 BG Breukelen, the Netherlands.

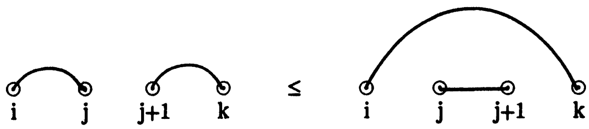


FIG. 1.1 *The inequality* $d[i,j]+d[j+1,k] \leq d[i,k]+d[j,j+1]$.

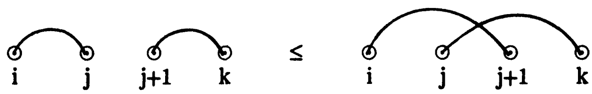


FIG. 1.2. *The inequality* $d[i,j]+d[j+1,k] \leq d[i,j+1]+d[j,k]$.

ϕ if $v > \text{MAX}\,\{\phi^{-1}(v), \phi(v)\}$ and a *valley* if $v < \text{MIN}\,\{\phi^{-1}(i), \phi(i)\}$. Let $P(\phi)$ denote the set of peaks, $V(\phi)$ the set of valleys, and

$$I(\phi) := N \backslash (P(\phi) \cup V(\phi))$$

the set of *intermediate points* of a permutation ϕ. Clearly, the number of peaks of a permutation is equal to the number of valleys, i.e., $|P(\phi)| = |V(\phi)|$ for every permutation ϕ. A permutation ϕ is a pyramidal tour if and only if it contains exactly one peak ($P(\phi) = \{n\}$) or, equivalently, exactly one valley ($V(\phi) = \{1\}$).

It is well known that the TSP is $\mathcal{NP}$-hard. However, for some special classes of distance matrices the TSP can be solved efficiently. For an excellent survey see [3]. One of these classes is

$$\mathbb{D}_{\text{PYR}} = \left\{ D \,\middle|\, \begin{array}{l} \text{For every Hamiltonian tour } \varphi \\ \text{there is a pyramidal tour } \rho \text{ such that } d(\rho) \leq d(\varphi) \end{array} \right\}.$$

For matrices in $\mathbb{D}_{\text{PYR}}$, [TSP] reduces to

$$\text{[PYR]} \quad \text{Minimize}\, \{d(\varphi) \,|\, \varphi \text{ is a pyramidal tour}\}.$$

This problem can be solved, using dynamic programming, in $\mathcal{O}(n^2)$ time (see [3, p. 100]). In [7] a class of (nonsymmetric) matrices has been identified such that the running time can be reduced to linear time. If $D \in \mathbb{D}_{\text{PYR}}$ then the associated TSP will be called *pyramidally solvable.*

In [1]–[9] several sets of conditions on distance matrices are introduced and it is shown that matrices satisfying these conditions belong to $\mathbb{D}_{\text{PYR}}$. This paper deals with classes of *symmetric* distance matrices ($d[i,j] = d[j,i]$ for all $i, j \in N$) that are subclasses of $\mathbb{D}_{\text{PYR}}$.

In §2 it will be shown that symmetric TSPs (STSPs) restricted to distance matrices in

$$\mathbb{D}_{\text{NEW}} \left\{ D = (d[i,j]) \,\middle|\, \begin{array}{c} d[i,j] = d[j,i] \text{ for all } i, j \\ d[i,j] + d[j+1,k] \leq d[i,k] + d[j,j+1] \\ \text{for all } i < j < j+1 < k \end{array} \right\}$$

are pyramidally solvable, i.e., that $\mathbb{D}_{\text{NEW}} \subset \mathbb{D}_{\text{PYR}}$. In §3 it is shown that $\mathbb{D}_{\text{NEW}} \not\subset \mathbb{D}_{\text{DEMI}}$ and $\mathbb{D}_{\text{DEMI}} \not\subset \mathbb{D}_{\text{NEW}}$ where $\mathbb{D}_{\text{DEMI}}$ is the class of symmetric Demidenko matrices, i.e.,

$$\mathbb{D}_{\text{DEMI}} \left\{ D = (d[i,j]) \,\middle|\, \begin{array}{c} d[i,j] = d[j,i] \text{ for all } i, j \\ d[i,j] + d[j+1,k] \leq d[i,j+1] + d[j,k] \\ \text{for all } i < j < j+1 < k \end{array} \right\}.$$

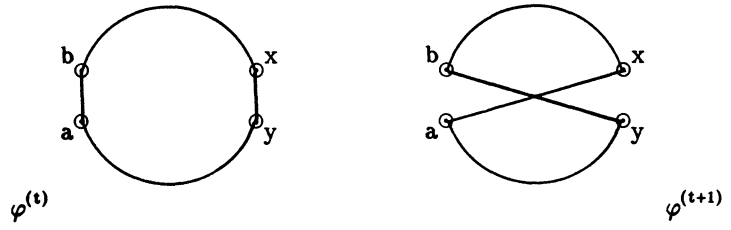


FIG. 2.1. *A 2-interchange.*

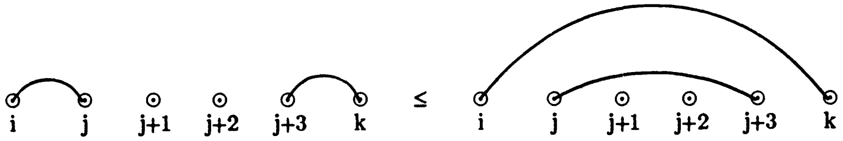


FIG. 2.2. *Inequality* (2) *with* $p = 3$.

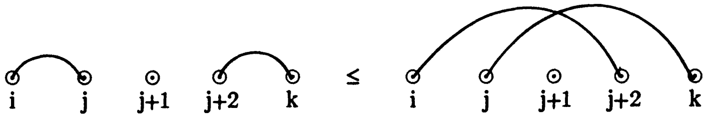


FIG. 2.3. *Inequality* (3) *with* $p = 2$.

This is relevant because until now $\mathbb{D}_{\text{DEMI}}$ was the most general class of symmetric matrices belonging in $\mathbb{D}_{\text{PYR}}$, meaning that all other subclasses of $\mathbb{D}_{\text{PYR}}$ known from the literature belong to $\mathbb{D}_{\text{DEMI}}$.

2. The STSP restricted to matrices in $\mathbb{D}_{\text{NEW}}$ is pyramidally solvable. In this section we will show that $\mathbb{D}_{\text{NEW}} \subset \mathbb{D}_{\text{PYR}}$, i.e., that if $D \in \mathbb{D}_{\text{NEW}}$ then for every Hamiltonian tour φ there is a pyramidal tour ρ such that $d(\rho) \leq d(\varphi)$. We will use an "improvement procedure." Starting from an arbitrary Hamiltonian cycle φ, a sequence $(\varphi^{(t)})_{t=1}^{T}$ (with $\rho^{(1)} := \varphi$) of Hamiltonian tours is constructed such that

$$d\left(\varphi^{(1)}\right) \geq d\left(\varphi^{(2)}\right) \geq \cdots \geq d\left(\varphi^{(T)}\right)$$

and T is the smallest index such that $\varphi^{(T)}$ is a pyramidal tour. The tour $\varphi^{(t+1)}$ is obtained from $\varphi^{(t)}$ by a 2-interchange, i.e., by replacing two edges of $\varphi^{(t)}$ by two new edges (see Fig. 2.1).

A 2-interchange is called *feasible* if and only if (1) the total length of the two new edges is no longer than the total length of the removed edges, and (2) the result of the 2-interchange is indeed a Hamiltonian tour, i.e., does not consist of two subtours. In order to characterize feasible 2-interchanges, consider the following lemma (see Figs. 2.2 and 2.3).

LEMMA 2.1. *Let* $D = (d[i,j]) \in \mathbb{D}_{\text{NEW}}$, *i.e.,* D *is a symmetric matrix such that*

$$(1) \qquad d[i,j] + d[j+1,k] \leq d[i,k] + d[j,j+1] \text{ for all } i < j < j+1 < k.$$

For all $i < j < j+1 < k$ *the following holds.*

(i) *For all odd integers* p *in* $\{1,2,\ldots,k-j-1\}$,

$$(2) \qquad d[i,j] + d[j+p,k] \leq d[i,k] + d[j,j+p].$$

(ii) *For all even integers* p *in* $\{0,1,2,\ldots,k-j-1\}$,

$$(3) \qquad d[i,j] + d[j+p,k] \leq d[i,j+p] + d[j,k].$$

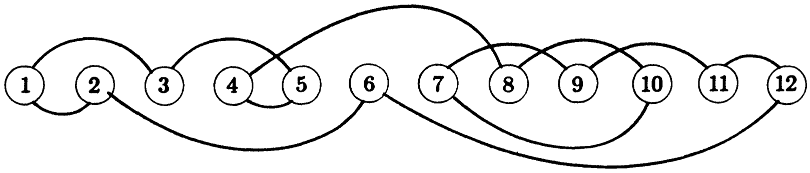


FIG. 2.4. *The Hamiltonian tour* $\varphi = (1,3,5,4,8,10,7,9,11,12,6,2)$.

Proof. Induction with respect to p will be used.

(i) For $p = 1$, (2) reduces to (1). Assume (2) holds for $p = P$ (where P is odd and at most $k - j - 3$), i.e.,

$$d[i,j] + d[j+P,k] \le d[i,k] + d[j,j+P]. \tag{4}$$

From (1) we have (with $j+P < j+P+1 < j+P+2 < k$)

$$d[j+P,j+P+1] + d[j+P+2,k] \le d[j+P,k] + d[j+P+1,j+P+2]. \tag{5}$$

Also from (1) (with $j < j+P < j+P+1 < j+P+2$),

$$d[j,j+P] + d[j+P+1,j+P+2] \le d[j,j+P+2] + d[j+P,j+P+1]. \tag{6}$$

Adding inequalities (4)–(6) and canceling

$$d[j,j+P] + d[j+P,k] + d[j+P,j+P+1] + d[j+P+1,j+P+2]$$

from both sides gives (2) for $p = P + 2$.

(ii) For $p = 0$ equality holds. Assume (3) holds for $p = P$ (where P is even and at most $k - j - 3$), i.e.,

$$d[i,j] + d[j+P,k] \le d[i,j+P] + d[j,k]. \tag{7}$$

From (1) (with $i < j+P < j+P+1 < j+P+2$) we have

$$d[i,j+P] + d[j+P+1,j+P+2] \le d[i,j+P+2] + d[j+P,j+P+1]. \tag{8}$$

Adding inequalities (7), (8), and (5) and canceling

$$d[i,j+P] + d[j+P,j+P+1] + d[j+P,k] + d[j+P+1,j+P+2]$$

from both sides gives (3) with $p = P + 2$. □

Let $F(\varphi) := \{[i,\varphi(i)] | \varphi(i) > i\}$ denote the set of *forward edges* and $B(\varphi) := \{[i,\varphi(i)] | \varphi(i) < i\}$ the set of *backward edges* of a tour φ. A pair of edges $([a,b];[x,y])$ is called a *transformable edge pair* (TEP) of the tour φ if and only if the following two conditions hold.

(1) $[a,b]$ and $[x,y]$ are in the same direction set of φ, meaning that either $[a,b]$ and $[x,y]$ are both in $F(\varphi)$ or $[a,b]$ and $[x,y]$ are both in $B(\varphi)$.

(2) Assuming $a = \text{MIN}\{a,b,x,y\}$ and $x = \text{MIN}\{x,y\}$, either $a < x < b < y$ and $b - x$ is even or $a < x < y < b$ and $y - x$ is odd.

By TEP(φ) we will denote the set of all transformable edge pairs of φ. A tour φ is said to contain a TEP if and only if TEP(φ) $\neq \emptyset$. The definition of a TEP is justified by the following lemma.

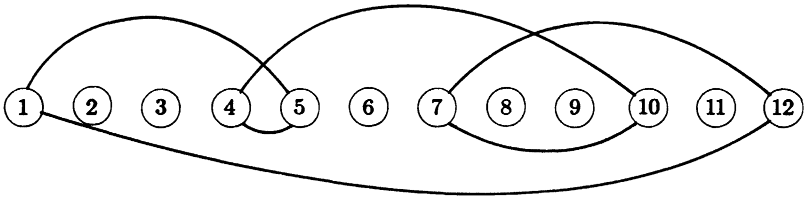


FIG. 2.5. *The subtour* $\pi_\varphi = (1,5,4,10,7,12)$.

LEMMA 2.2. *If the tour* φ *contains a* TEP *then there is a feasible* 2*-interchange on* φ.

Proof. Let $([a,b];[x,y]) \in \mathrm{TEP}(\varphi)$ and φ' the tour that arises from φ by replacing $[a,b]$ and $[x,y]$ by $[a,x]$ and $[b,y]$ and reversing the path from b to x. From Lemma 2.1 it follows that $d[a,b]+d[x,y] \geq d[a,x]+d[b,y]$. Since D is symmetric it follows that $d(\varphi') \leq d(\varphi)$. Because $[a,b]$ and $[x,y]$ are in the same direction set, φ' is a Hamiltonian tour. □

In order to use the above-described procedure it has to be shown that a Hamiltonian tour φ contains a TEP if and only if it is not pyramidal. To that end the following concepts are introduced. A path $[i_1,\ldots,i_k]$ is an *ordered path* in a Hamiltonian tour φ if and only if

(1) $i_t = \varphi(i_{t-1})$ for $t = 2,\ldots,k$, and

(2) all edges $[i_{t-1},i_t](t=2,\ldots,k)$ are in the same direction set of φ.

A Hamiltonian tour φ can be divided into $2|P(\varphi)|$ ordered paths as follows. There are $|P(\varphi)|$ ordered paths with all edges in one direction set starting at a vertex in $V(\varphi)$ and ending at a vertex in $P(\varphi)$, and $|P(\varphi)|$ ordered paths with all edges in the other direction set starting at a vertex in $P(\varphi)$ and ending at a vertex in $V(\varphi)$.

Example. Let $\varphi = (1,3,5,4,8,10,7,9,11,12,6,2)$. See Fig. 2.4. So, $V(\varphi) = \{1,4,7\}, P(\varphi) = \{5,10,12\}, I(\varphi) = \{2,3,6,8,9,11\}, F(\varphi) = \{[1,3];[3,5];[4,8];[8,10];[7,9];[9,11];[11,12]\}$, and $B(\varphi) = \{[5,4];[10,7];[12,6];[6,2];[2,1]\}$. The $2|P(\varphi)| = 6$ ordered paths in φ are $[1,3,5];[4,8,10];[7,9,11,12]$ with edges in $F(\varphi)$ and $[5,4]$; $[10,7];[12,6,2,1]$ with edges in $B(\varphi)$. □

Let π_φ be the (sub)tour that arises from a Hamiltonian tour φ by deleting all intermediate points, i.e., the $2|P(\varphi)|$ ordered paths in φ are replaced by edges with the same endpoints as the ordered paths. For the φ in the example above this gives $\pi_\varphi = (1,5,4,10,7,12)$. See Fig. 2.5.

LEMMA 2.3. *For any Hamiltonian tour* φ, *let* $[a,b]$ *and* $[x,y]$ *be any two edges in the same direction set of* π_φ *such that* $a < x < b$ *and let* $[a = i_0,i_1,i_2,\ldots,i_{k-1}, i_k = b](k \geq 1)$ *and* $[x = j_0,j_1,j_2,\ldots,j_{l-1},j_l = y](l \geq 1)$ *be the two corresponding ordered paths in* φ. *The following holds. If* $([a,b];[x,y]) \in \mathrm{TEP}(\pi_\varphi)$ *then there is an* $r \in \{0,\ldots,k-1\}$ *and an* $s \in \{0,\ldots,l-1\}$ *such that* $([i_r,i_{r+1}];[j_s,j_{s+1}]) \in \mathrm{TEP}(\varphi)$.

Proof. According to the definition of a TEP we distinguish between two cases.

(1) Assume $a < x < y < b$. Since $y-x$ is odd, there is an $s \in \{0,\ldots,l-1\}$ such that $j_{s+1}-j_s$ is odd. Define $u \in \{0,\ldots,k-1\}$ such that i_u is the largest i smaller than j_s and $v \in \{0,\ldots,k-1\}$ such that i_v is the smallest i larger than j_{s+1}. If $v = u+1$ then $([i_u,i_{u+1}];[j_s,j_{s+1}]) \in \mathrm{TEP}(\varphi)$ because $i_u < j_s < j_{s+1} < i_{u+1}$ and $j_{s+1}-j_s$ is odd. If $v > u+1$ then we have $i_u < j_s < i_{u+1} \leq i_{v-1} < j_{s+1} < i_v$. If $i_{u+1}-j_s$ is even then $([i_u,i_{u+1}];[j_s,j_{s+1}]) \in \mathrm{TEP}(\varphi)$ because $i_u < j_s < i_{u+1} < j_{s+1}$ and $i_{u+1}-j_s$ is even. If $j_{s+1}-i_{v-1}$ is even then $([i_{v-1},i_v];[j_s,j_{s+1}]) \in \mathrm{TEP}(\varphi)$ because $j_s < i_{v-1} < j_{s+1} < i_v$ and $j_{s+1}-i_{v-1}$ is even. If both $i_{u+1}-j_s$ and $j_{s+1}-i_{v-1}$ are odd then $i_{v-1}-i_{u+1}$ is odd. So there is a $w \in \{u+1,\ldots,v-2\}$ such that $i_{w+1}-i_w$ is odd. In this case $([i_w,i_{w+1}];[j_s,j_{s+1}]) \in \mathrm{TEP}(\varphi)$ because $j_s < i_w < i_{w+1} < j_{s+1}$ and $i_{w+1}-i_w$ is odd.

(2) Assume $a < x < b < y$. Without loss of generality we may assume that $a =$

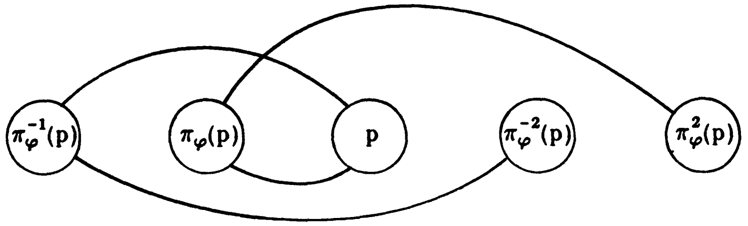


FIG. 2.6. *Part of the tour π_φ.*

$x-1$ (this follows from the fact that $([a,b];[x,y])$ is a TEP if and only if $([x-1,b];[x,y])$ is a TEP). Since $b-x$ is even, $b-a(=b-x+1)$ is odd. So there is an $r \in \{0,\dots,k-1\}$ such that $i_{r+1}-i_r$ is odd. In a similar way as in the first case it is possible to find an $s \in \{0,\dots,l-1\}$ such that $([i_r,i_{r+1}];[j_s,j_{s+1}]) \in \mathrm{TEP}(\varphi)$. □

Note that the opposite of Lemma 2.3 does not always hold. For instance the ordered paths $[1,4,7]$ and $[2,6]$ (with edges in the same direction set) contain a TEP; $(([1,4];[2,6])$ is a TEP and so is $([2,6];[4,7]))$ but $([1,7];[2,6])$ is not a TEP.

LEMMA 2.4. $\mathrm{TEP}(\varphi)=\emptyset$ *if and only if φ is a pyramidal tour.*

Proof. If φ is a pyramidal tour then it does not contain a TEP because there are no two edges $[a,b]$ and $[x,y]$ in φ such that these are in the same direction class and $a<x<b$.

Assume φ is not pyramidal. It has to be shown that φ contains a TEP. From Lemma 2.3 it follows that if π_φ contains a TEP then φ contains a TEP. So it suffices to show that π_φ contains a TEP. Let p be the *smallest* peak of φ (and hence of π_φ). Consider the vertices $\pi_\varphi^{-2}(p), \pi_\varphi^{-1}(p), \pi_\varphi(p)$, and $\pi_\varphi^2(p)$. Note that $\pi_\varphi^{-1}(p), \pi_\varphi(p) \in V(\varphi)$ and $\pi_\varphi^{-2}(p), \pi_\varphi^2(p) \in P(\varphi)$. Furthermore, $\mathrm{MAX}\{\pi_\varphi^{-1}(p),\pi_\varphi(p)\} < p$ (p is a peak), and $\mathrm{MIN}\{\pi_\varphi^{-2}(p),\pi_\varphi^2(p)\} > p$ because $\pi_\varphi^{-2}(p)$ and $\pi_\varphi^2(p)$ are both (not necessarily different) peaks, p is the smallest peak and (since φ is not pyramidal) $p<n$. The following holds: either $([\pi_\varphi^{-1}(p),p];[\pi_\varphi(p),\pi_\varphi^2(p)]) \in \mathrm{TEP}(\pi_\varphi)$ or $([\pi_\varphi^{-2}(p),\pi_\varphi^{-1}(p)];[p,\pi_\varphi(p)]) \in \mathrm{TEP}(\pi_\varphi)$.

It is obvious that $[\pi_\varphi^{-1}(p),p]$ and $[\pi_\varphi(p),\pi_\varphi^2(p)]$ are in the same direction class and that $[\pi_\varphi^{-2}(p),\pi_\varphi^{-1}(p)]$ and $[p,\pi_\varphi(p)]$ are both in the other direction class. Which of the two pairs is a TEP depends on

(i) whether $\pi_\varphi^{-1}(p) < \pi_\varphi(p)$ or $\pi_\varphi^{-1}(p) > \pi_\varphi(p)$ and

(ii) whether $p-\mathrm{MAX}\{\pi_\varphi^{-1}(p),\pi_\varphi(p)\}$ is odd or even.

For instance when $\pi_\varphi^{-1}(p) < \pi_\varphi(p) < p < \pi_\varphi^{-2}(p) < \pi_\varphi^2(p)$ and $p-\pi_\varphi(p)$ is odd then $([\pi_\varphi^{-2}(p),\pi_\varphi^{-1}(p)];[p,\pi_\varphi(p)]) \in \mathrm{TEP}(\pi_\varphi)$. See Fig. 2.6. The other cases are left to the reader. □

The improvement procedure can now be described more formally.

```
Procedure IMPROVE
Input:    A Hamiltonian tour $\varphi$.
Output:   A pyramidal tour.
Step 0    [Initialization]
          Set $\varphi^{(1)} := \varphi$ and $t = 1$. Go to Step 1.
Step 1    If TEP $(\varphi^{(t)}) = \emptyset$ then stop: $\varphi^{(t)}$ is a pyramidal tour.
          Else find $([a,b];\ [x,y]) \in$ TEP $(\varphi^{(t)})$ and go to Step 2.
Step 2    Obtain $\varphi^{(t+1)}$ from $\varphi^{(t)}$ by replacing $[a,b]$ and $[c,d]$ by $[a,x]$
          and $[b,y]$ and reversing the path from $b$ to $y$.
          Return to Step 1 with $t := t + 1$.
```

Lemma 2.4 assures the existence of a TEP until a pyramidal tour is found. By

Lemma 2.2 the 2-interchange in Step 2 is feasible. It remains to show that procedure IMPROVE ends after a finite number of steps for any initial Hamiltonian tour. Intuitively, this means that we have to show that $\varphi^{(t+1)}$ is "more pyramidal" than $\varphi^{(t)}$.

Define the *pyramidality number* $K(\varphi)$ of a Hamiltonian tour φ as follows.

$$K(\varphi) := \sum_{p \in P(\varphi)} p - \sum_{v \in V(\varphi)} v.$$

With the φ in the example above, $K(\varphi) = (5 + 10 + 12) - (1 + 4 + 7) = 15$. For every Hamiltonian tour it holds that $\lfloor n/2 \rfloor \lceil n/2 \rceil \geq K(\varphi) \geq n - 1$. The upper bound is tight for Hamiltonian tours of the type

$$(1, n, 2, n-1, 3, n-2, \cdots).$$

Furthermore, $K(\varphi) = n - 1$ if and only if φ is a pyramidal tour. So, the lower $K(\varphi)$ the "more pyramidal" φ is. An alternative way to define $K(\varphi)$ is as follows. Let $k_\varphi(i), i = 1, \ldots, n-1$, denote the number of edges $[u, v]$ of φ such that $\text{MIN}\{u, v\} \leq i < i + 1 \leq \text{MAX}\{u, v\}$. Note that $k_\varphi(i)$ is even and ≥ 2 for all i. It follows that

$$K(\varphi) = \frac{1}{2} \sum_{i=1}^{n-1} k_\varphi(i).$$

For instance with the example above we have $k_\varphi(1) = k_\varphi(2) = k_\varphi(3) = k_\varphi(5) = k_\varphi(6) = k_\varphi(10) = k_\varphi(11) = 2$ and $k_\varphi(4) = k_\varphi(7) = k_\varphi(8) = k_\varphi(9) = 4$, so $K(\varphi) = 15$.

LEMMA 2.5. *Let $\{\varphi^{(t)}\}$ be a sequence produced by procedure* IMPROVE. *If $\varphi^{(t)}$ is not a pyramidal tour then $K(\varphi^{(t+1)}) < K(\varphi^{(t)})$.*

Proof. In Step 2 of procedure IMPROVE $[a, b]$ and $[x, y]$ are replaced by $[a, x]$ and $[b, y]$. So $k_{\varphi^{(t+1)}}(i) = k_{\varphi^{(t)}}(i) - 1$ for $i = x, \ldots, \text{MIN}\{y, b\} - 1$, hence

$$K\left(\varphi^{(t+1)}\right) = K\left(\varphi^{(t)}\right) - (\text{MIN}\{b, y\} - x),$$

or equivalently

$$K\left(\varphi^{(t+1)}\right) - K\left(\varphi^{(t)}\right) = x - \text{MIN}\{b, y\} < 0. \qquad \square$$

By Lemma 2.5, $K(\varphi^{(T)}) = n - 1$ for some finite integer T and $\varphi^{(T)}$ is a pyramidal tour. So procedure IMPROVE ends after a finite number of steps. The results in this section are summarized in the following theorem.

THEOREM 2.6. $\mathbb{D}_{\text{NEW}} \subset \mathbb{D}_{\text{PYR}}$.

3. $\mathbb{D}_{\text{NEW}}$ is not a subclass of $\mathbb{D}_{\text{DEMI}}$.

As mentioned in §1, A matrix D is a *symmetric Demidenko matrix* ($D \in \mathbb{D}_{\text{DEMI}}$) if and only if

$$d[i, j] + d[j+1, k] \leq d[i, j+1] + d[j, k] \quad \text{for all } i < j < j+1 < k.$$

This condition is the symmetric equivalent of the conditions introduced in [2] (see also [3, p. 102]). In [9] a simple proof is given for $\mathbb{D}_{\text{DEMI}} \subset \mathbb{D}_{\text{PYR}}$. Until now $\mathbb{D}_{\text{DEMI}}$ was the most general class of matrices belonging to $\mathbb{D}_{\text{PYR}}$, meaning that all other subclasses of $\mathbb{D}_{\text{PYR}}$ known from the literature belong to $\mathbb{D}_{\text{DEMI}}$. The purpose of this section is to show that $\mathbb{D}_{\text{DEMI}} \not\subset \mathbb{D}_{\text{NEW}}$ and $\mathbb{D}_{\text{NEW}} \not\subset \mathbb{D}_{\text{DEMI}}$.

Matrices in $\mathbb{D}_{\mathrm{NEW}}$ can be constructed in a similar way as was shown in [9] for matrices in $\mathbb{D}_{\mathrm{DEMI}}$. It is easy to construct matrices that are in $\mathbb{D}_{\mathrm{NEW}}$ but not in $\mathbb{D}_{\mathrm{DEMI}}$ or matrices that are in $\mathbb{D}_{\mathrm{DEMI}}$ but not in $\mathbb{D}_{\mathrm{NEW}}$ or any value of n. For details we refer to [8]. Here we give an example for $n = 4$.

Example. Consider the following two matrices.

$$D_1 = \begin{bmatrix} 0 & 4 & 2 & 4 \\ 4 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 \\ 4 & 1 & 0 & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 & 4 & 4 & 2 \\ 4 & 0 & 1 & 0 \\ 4 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{bmatrix}.$$

It is easy to check that $D_1 \in \mathbb{D}_{\mathrm{NEW}}$ and $D_1 \notin \mathbb{D}_{\mathrm{DEMI}}$ and that $D_2 \in \mathbb{D}_{\mathrm{DEMI}}$ and $D_2 \notin \mathbb{D}_{\mathrm{NEW}}$. □

Acknowledgments. A substantial part of the research for this paper was done while the author worked at the department of Econometrics of the University of Groningen, the Netherlands. An (anonymous) referee gave the upper bound on $K(\varphi)$ in §2 and other useful remarks.

REFERENCES

[1] R. E. Burkard, J. A. A. van der Veen, *Universal conditions for algebraic traveling salesman problems to be efficiently solvable*, Optimization, 22 (1991), pp. 787–814.

[2] V. M. Demidenko, *The traveling salesman problem with asymmetric matrices*, Vesti Akad. Navuk BSSR, Ser. Fiz.-Mat. Navuk, 1 (1979), pp. 29–35. (In Russian.)

[3] R. C. Gilmore, E. L. Lawler, and D. B. Shmoys, *Well-solved special cases*, in The Traveling Salesman Problem, E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, eds., Wiley, Chichester, 1985, pp. 87–143.

[4] K. Kalmanson, *Edgeconvex circuits and the traveling salesman problem*, Canad. J. Math., 27 (1975), pp. 1000–1010.

[5] P. S. Klyaus, *Structure of the optimal solution of certain classes of traveling salesman problems*, Vesti Akad. Navuk BSSR, Ser. Fiz.-Mat. Navuk, 6 (1976), pp. 95–98. (In Russian.)

[6] F. Supnick, *Extreme Hamiltonian lines*, Ann. Math., 65 (1957), pp. 179–201.

[7] J. K. Park, *A special case of the n-vertex traveling salesman problem that can be solved in $\mathcal{O}(n)$ time*, Inform. Process. Lett., 40 (1991), pp. 247–254.

[8] J. A. A. van der Veen, *Solvable Cases of the Traveling Salesman Problem with Various Objective Functions*, Ph.D. thesis, University of Groningen, 1992.

[9] J. A. A. van der Veen, G. Sierksma, and R. van Dal, *Pyramidal tours and the traveling salesman problem*, European J. Oper. Res., 52 (1991), pp. 90–102.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 593–597, November 1994

007

DRAWING GRAPHS ON SURFACES *

ARJANA ŽITNIK†

Abstract. Every graph that is 2-cell embedded in a closed orientable surface can be drawn in some fundamental polygon of the surface so that the boundary of the polygon consists of edges and vertices of the graph; it can also be drawn so that all the vertices of the graph are inside the polygon and no edges cross the boundary of the polygon more than once. A necessary and sufficient condition is found to determine whether or not a given embedding of a graph in a surface has one or the other representation in the standard polygon of the surface, the one of the form $a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\ldots$. This leads to a polynomial-time algorithm, assuming that the surface is fixed.

Key words. graphs, embeddings, drawing graphs, polygons, algorithms

AMS subject classifications. 05C10, 05C75, 05C85, 68R10

1. Introduction.

The definitions used in this paper are more or less standard and can be found in [2], [3], [7]. We assume graphs to be finite, connected, and undirected. A surface will always stand for an orientable surface and we denote the oriented surface with $k > 0$ handles by S_k.

A standard method of depicting a graph embedded on a surface is to draw the graph on the fundamental polygon of the surface, e.g., a polygon in the plane with its sides oriented and labeled to indicate how the boundary of the polygon can be identified to recover the surface. Of course, all the edges cannot lie inside the polygon. Some of them may lie on the boundary of the polygon; some may cross over the boundary of the polygon. The problem is that some edges may repeatedly cross over the polygon's boundary, making the drawing unclear and difficult to visually verify the desired adjacencies.

We avoid such problems in the following two cases.

- The boundary of the polygon consists exclusively of the vertices and edges of the graph. We say that the graph is *compounded* in the polygon (see Fig. 1).
- All the vertices of the graph are inside the polygon and each edge is allowed to cross the polygon's boundary once at most. If a graph G is drawn in a polygon P in this way, we say that G is *drawn nicely* in P.

Every embedding of a graph G can in both cases be depicted in some fundamental polygons of the surface. In the first case we paste together all the faces of the graph G in some way to obtain a single polygon. In the second case we paste together all the faces of the dual graph and then draw G in this polygon—the vertices of G inside the faces of G^* and the edges of G cross the edges of G^* in the middle.

We can represent a surface of genus k with a polygon in the normal form, the polygon with edges identified in order $a_1b_1a_1^{-1}b_1^{-1}\cdots a_kb_ka_k^{-1}b_k^{-1}$. We call this polygon the *standard polygon* of the surface S_k and denote it by P_k. The number of different polygons representing the same surface increases with the genus of the surface. Some of them may have very many edges on the boundary. For this reason we consider only the drawings of graphs in standard polygons. Unfortunately, not every graph with a

* Received by the editors December 28, 1992; accepted for publication (in revised form) October 29, 1993.

† Department of Theoretical Computer Science, Institute of Mathematics, Physics, and Mechanics, University of Ljubljana, Jadranska 19, 61000 Ljubljana, Slovenia.

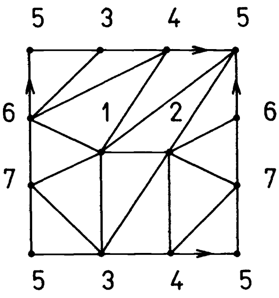


FIG. 1. *A drawing of K_7 in the polygon P_1 which represents the torus.*

given embedding can be drawn nicely or compounded in the standard polygon of the surface. We first give the rather obvious condition whether a 2-cell embedded graph can be compounded in the standard polygon of the surface. As the main point of this article we present a necessary and sufficient condition for a graph with a given embedding to be drawn nicely in the standard polygon of the surface. For every surface of fixed genus this leads to a polynomial-time algorithm, which is at present practically useful only for the torus.

2. Main result.

Throughout this section G will denote a graph that is 2-cell embedded on the surface S_k.

Let B be a graph embedded on the oriented surface S_k. We call it a *bouquet of k crossing cycles* or a k-bouquet, if it consists of $2k$ cycles, intersecting at one vertex, say v, and the cyclic order of edges around v is the following: $e_{1,1}, e_{2,1}, e_{1,2}, e_{1,2}, e_{2,2}, \dots,$ $e_{2k-1,1}, e_{2k,1}, e_{2k-1,2}, e_{2,k,2}$. Here $e_{i,1}$ and $e_{i,2}$ denote the edges of the ith cycle which have v as an endvertex ($e_{i,1}$ and $e_{i,2}$ may be a single edge, a loop).

If a graph can be compounded in the standard polygon P_k, it contains a k-bouquet, which is the boundary of the polygon. The converse also holds: if a graph embedded in S_k contains a k-bouquet B_k, we can cut S_k along the edges of B_k and obtain the polygon P_k. This proves the following lemma.

LEMMA 2.1. *Let a graph G be embedded in $S_k, k > 0$. Then G can be compounded in the standard polygon P_k if and only if it contains a k-bouquet as a subgraph.*

COROLLARY 2.2. *If a graph G does not contain a vertex of degree $4k$ or more, it cannot be compounded in P_k.*

The *dual graph* of a 2-cell embedded graph G (relative to this embedding) is a graph G^* embedded on S_k such that the vertices of G^* correspond to the faces of G and every edge of G^* corresponds to precisely one edge of G. We denote the edge corresponding to an edge $e \in E(G)$ by e^*. Two vertices of G^* are adjacent if and only if the boundaries of their corresponding faces of G share a common edge. G^* is embedded in the following way: on each edge of G we choose a point—it is called the *midpoint* of the edge. We put a vertex of G^* in each face of G and connect it to all the midpoints on the boundary of the face.

LEMMA 2.3. *We can draw a graph G on the standard polygon P_k nicely with no faces split in more than two parts, except the one in the corner of the polygon, which is split in $4k$ parts, if and only if the dual graph of G can be compounded in the standard polygon P_k.*

Since the proof is obvious, we omit it.

There exist embedded graphs that can be drawn nicely on the standard polygon, but their dual graphs cannot be compounded in it. The following example will show this. Let K_7' be the embedded graph K_7 from Fig. 2; we only add a vertex on the edge e that is cut twice. We get a nice embedding of K_7'. The dual graph of K_7 cannot

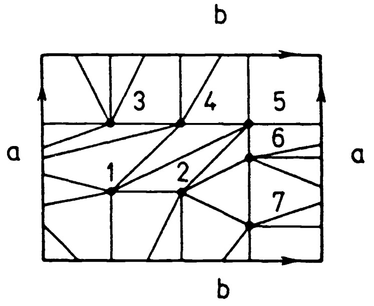


FIG. 2. *An alternative drawing of K_7 in P_1.*

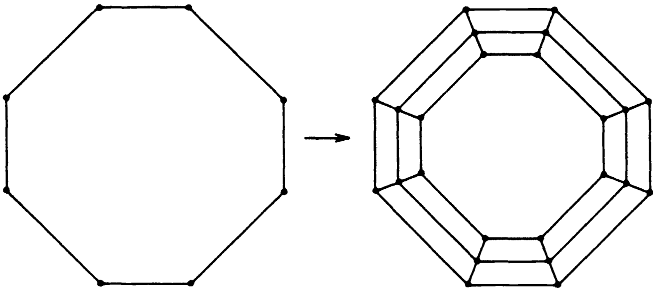


FIG. 3. *The rule for the construction of the associated graph.*

be compounded in P_1, since all its vertices are of degree 3. We get the graph $K_7'^*$ from K_7^* by only adding a parallel edge to e^* which is embedded in the same way. So we cannot find a 1-bouquet in $K_7'^*$ and therefore $K_7'^*$ cannot be compounded in the standard polygon P_1.

Let $f_1, \dots, f_j$ be the faces of the embedded graph G, with d_i edges on the boundary for $i = 1 \cdots j$. We construct another graph that is embedded in S_k by putting a "net" with $\max\{\lfloor d_i/4 \rfloor, \lfloor (d_i - 4k)/2 \rfloor\}$ layers and d_i sides in each face $f_i, i = 1 \cdots j$, of graph G, so the boundary of the face is the 0th layer (see Fig. 3). We call this (embedded) graph the *associated graph of graph G*. We denote it by $\tilde{G}$. The associated graph is obviously simple, provided that the initial graph does not have any vertices of degree one.

THEOREM 2.4. *We can draw a graph G on the standard polygon P_k nicely if and only if the dual graph $\tilde{G}^*$ of the associated graph can be compounded in the standard polygon P_k.*

Proof. If $\tilde{G}^*$ can be compounded in P_k, we can draw $\tilde{G}$ on the polygon P_k nicely by Lemma 2.3. Since G is a subgraph of $\tilde{G}$, it can be drawn on the polygon in the same way; the superfluous edges simply are omitted.

The other way is a little more complicated. First, let all the faces of G be simple; that means that no edge occurs on the boundary of the same face twice. Each edge of G corresponds to a layer of vertices of $\tilde{G}^*$ in two faces of G. We denote these vertices by triples (e, f, i), where e is the label of the edge, f is the label of the face, and i is the number of the layer. The layer that is nearest to the boundary of the face of G has number 1. We denote the vertex of $\tilde{G}^*$ in the middle of the face f simply by (f). Suppose G is drawn nicely on P_k with symbolic description $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_k b_k a_k^{-1} b_k^{-1}$. We can construct a k-bouquet in $\tilde{G}^*$ using the drawing of G in P_k. Let f_1 be the face of G by the corner of the polygon; let $e_1, e_2, \dots, e_{j1}$ be the cut edges; and let $f_2, f_3, \dots, f_{j1}$ be the cut faces in the same order as they appear along the side a_1 of the polygon. All the edges e_i are different, but some faces may be the same.

Some new notation is now needed. Let d_i be the number of edges on the boundary of f_i and $l_i = \max\{\lfloor d_i/4 \rfloor, \lfloor (d_i - 4k)/2 \rfloor\}$ the number of layers in the face. Let w_i be a segment of the boundary of f_{i+1} that contains the edges e_i and e_{i+1}. There are two such paths, we take the shorter one. With s_i' we denote the number of edges between e_i and e_{i+1} on w_i, including e_i and e_{i+1}. Let $s_i = \lfloor s_i'/2 \rfloor$ denote the layer we use to come from the vertex $(e_i, f_{i+1}, 1)$ of the graph $\tilde{G}^*$ to the vertex $(e_{i+1}, f_{i+1}, 1)$ in the face f_{i+1} of the graph G. Let e_i^+ be the edge that follows e_i on w_i and let e_{i+1}^- be the edge before e_{i+1}.

Now we can construct one of the cycles, say C_1, of the k-bouquet. Since $\tilde{G}^*$ is a simple graph, we can represent C_1 by a sequence of vertices as follows:

$$\begin{aligned}
&(f_1) - (e_1, f_1, l_1) - (e_1, f_1, l_1 - 1) - \cdots - (e_1, f_1, 1) - (e_1, f_2, 1) \\
&\quad - (e_1, f_2, 2) - \cdots - (e_1, f_2, s_1) - (e_1^+, f_2, s_1) - (e_1^{++}, f_2, s_1) - \cdots - (e_2^-, f_2, s_1) \\
&\quad - (e_2, f_2, s_1) - (e_2, f_2, s_1 - 1) - \cdots - (e_2, f_2, 1) - (e_2, f_3, 1) - \cdots - (e_j, f_j, 1) \\
&\quad - (e_j, f_1, 1) - (e_j, f_1, 2) - \cdots - (e_j, f_1, l_1) - (f_1).
\end{aligned}$$

We construct the other $2k - 1$ cycles similarly. If some faces are not simple, we need an additional label to tell us on which side of the edge a layer of vertices of $\tilde{G}^*$ belongs. If e is an edge that appears on the boundary of the face f twice, we represent the "cut" by an edge $(e, f, 1, side1) - (e, f, 1, side2)$ in the cycle. If G contains vertices of degree one, $\tilde{G}^*$ is not simple. Let v be a vertex of degree one that is connected to G with an edge e. Then $\tilde{G}^*$ has parallel edges between the vertices $(e, f, 1, side1)$ and $(e, f, 1, side2)$. Since it does not make any difference which one of these two edges we take, we will not make distinction between them. We construct the cycles as before. The cycles obviously intersect at (f_1) in the right way.

Suppose we cut the surface S along some cycle in such a way that we cut no edges of G more than once. Let $u_1, \ldots, u_m$ be the boundary walk of some face, which is cut. Say that we cut it from the edge u_1 to the edge u_i. Then we cannot cut this face from edge u_k to u_l, if $1 < k < i$ and $i < l < m$, because we would have to cross the first cut and then we would not have a cycle. That implies that $C_i, i = 1, \ldots, 2k$ really are cycles and that they intersect only at vertex (f_1). □

An algorithm for determining whether or not a graph which is 2-cell embedded in S_k can be nicely drawn in the standard polygon P_k is derived easily. We have to construct the graph $\tilde{G}^*$ and find a k-bouquet in it. If a k-bouquet does not exist, then G cannot be drawn in P_k nicely. Otherwise G can be drawn nicely in P_k and, having the k-bouquet, we can actually draw G in the polygon, using some polynomial-time algorithm for drawing plane graphs, for example [1], [6].

We reduce the problem of how to find a k-bouquet to the k disjoint paths problem, which is the following: given a graph G and pairs $(s_1, t_1), \ldots, (s_k, t_k)$ of vertices, are there k paths joining the corresponding pairs of vertices (s_i with t_i) that are vertex disjoint? Robertson and Seymour proved (see [4]) that for every fixed k there exists a polynomial-time algorithm for the k disjoint paths problem. Its running time is $O(|V(G)|^3)$.

Using the algorithm of Robertson and Seymour, we can find a k-bouquet in $\tilde{G}^*$ in polynomial time regarding the number of vertices and edges of G. All the other steps are obviously polynomial. So if the genus of the surface is fixed, the algorithm is polynomial-time. A similar algorithm can be derived for compounding graphs in the standard polygon of the surface that has to find a k-bouquet in the graph G.

We have to mention that the k disjoint paths algorithm of Robertson and Seymour is not useful in practice, since it involves enormous constants. So far there are efficient algorithms for finding two disjoint paths, for example the one of Y. Shiloach [5], which needs $O(n^2)$ time. So the algorithm can be efficiently implemented at least for the torus, using the algorithm of [5] to find two disjoint paths.

3. Some other problems. If a 2-cell embedded graph G can be drawn nicely in the standard polygon of the surface it is *embedded nicely* on the surface and we call the embedding of G *nice*. We have only considered the problem of whether a given embedding of a graph is nice. Another problem is how to embed a graph nicely on a surface. There are graphs that can be 2-cell embedded in a surface S, however, none of these embeddings is nice. An example of such graph is K_7, which can be 2-cell embedded in the torus, but it does not have a nice embedding in it. The question is if for every graph there exists a surface in which G can be 2-cell embedded nicely.

Acknowledgment. This work is a part of my Diploma Thesis at the University of Ljubljana. I wish to thank my advisor Professor Tomaž Pisanski for pointing out the problem and for many helpful comments.

REFERENCES

[1] N. Chiba, K. Onoguchi, and T. Nishizeki, *Drawing plane graphs nicely*, Acta Inform., 22 (1985), pp. 187–201.

[2] J. L. Gross and T. W. Tucker, *Topological Graph Theory*, John Wiley, New York, 1987.

[3] G. Ringel, *Map Color Theorem*, Springer-Verlag, Berlin, 1984.

[4] N. Robertson and P. D. Seymour, Graph minors XIII. The disjoint paths problem, manuscript, 1986.

[5] Y. Shiloach, *A polynomial solution to the undirected two paths problem*, J. Assoc. Comput. Mach., 27 (1980), pp. 445–456.

[6] W. T. Tutte, *How to draw a graph*, Proc. London Math. Soc. (3), 13 (1963), pp. 743–768.

[7] A. T. White, *Graphs, Groups and Surfaces*, North-Holland, Amsterdam, 1984.

ON UNIVERSAL CYCLES FOR k-SUBSETS OF AN n-SET*

GLENN HURLBERT[†]

Abstract. A *universal cycle,* or *Ucycle,* for k-subsets of $[n] = \{1, \ldots, n\}$ is a cyclic sequence of $\binom{n}{k}$ integers with the property that each subset of $[n]$ of size k appears exactly once consecutively in the sequence. Chung, Diaconis, and Graham have conjectured their existence for fixed k and large n when $n | \binom{n}{k}$. Here the Ucycles for $k = 3$, 4, 6 and large n relatively prime to k are exhibited.

Key words. universal cycle, de Bruijn cycle

AMS subject classifications. 05B30, 05C38, 05C45, 94A55

1. Introduction. In 1894 the following binary sequence was discovered by Flye-Sainte Marie [M]:

$$00011101.$$

The interesting property of this sequence is that, when regarded as a cyclic sequence, every binary triple appears uniquely as a consecutive block of digits. Indeed, 000, 001, 011, 111, 110, 101, 010, 100 is the corresponding list of binary triples. It was also discovered that for all n and k analogous k-ary sequences exist listing k-ary n-tuples uniquely. This result went largely unnoticed until the sequences were rediscovered in 1946 by de Bruijn [B] and Good [G], and they have come to be known as *de Bruijn cycles.* For interesting results, generalizations, applications, and history see [F], [H].

Recently, Chung, Diaconis, and Graham [C] have generalized such sequences so as to list combinatorial families other than k-ary n-tuples including permutations of $[n]$, partitions of $[n]$, k-subsets of $[n]$, and k-dimensional subspaces of an n-dimensional vector space over a finite field. They call such sequences *universal cycles,* or *Ucycles.*

In this paper we discuss only the case of Ucycles for k-subsets of $[n]$. An example with $k = 3, n = 8$ is as follows.

$$1356725 \quad 6823472 \quad 3578147 \quad 8245614 \quad 5712361 \quad 2467836 \quad 7134583 \quad 4681258$$

The list of subsets begins with $135, 356, 567, \ldots,$ and ends with $\ldots, 258, 581, 813$. (Often, we will leave out the brackets in set notation to improve readability.) And, of course, there are $\binom{8}{3} = 56$ digits in the sequence, one for each subset. In the example above, spaces were used only to help the reader notice that each block is just an additive translation of the previous block. That is, we add 5 modulo 8 to the digits of one block to obtain the digits of the next block. This will be an important feature of Ucycle construction.

We first remark that any ordering of $[n]$ is a Ucycle both for 1- and $(n-1)$-subsets. These we call trivial Ucycles. Also, any listing of the vertices of an Eulerian circuit in the complete graph K_n (n *odd*) is a Ucycle for 2-subsets of $[n]$. Thus, we shall always assume that $k \geq 3$ and $n \geq k + 2$.

Next, we notice that each integer must occur equally often in a Ucycle since each integer is in equally many k-subsets of $[n]$. So for the existence of such Ucycles we

* Received by the editors August 6, 1991; accepted for publication November 8, 1993.

[†] Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804.

have the following necessary condition

$$\text{(NC)} \qquad n \text{ divides } \binom{n}{k}.$$

Equivalently, we may write that k divides $\binom{n-1}{k-1}$. This carries the interpretation that any digit in the Ucycle is in k k-subsets and any integer is in $\binom{n-1}{k-1}$ k-subsets, so $\binom{n-1}{k-1}/k$ is the number of occurrences of each integer in the Ucycle. Notice that (NC) holds whenever n is relatively prime to k. Indeed, $\binom{n}{k} = n\binom{n-1}{k-1}/k$ is an integer and n and k have no common factors, so $\binom{n-1}{k-1}/k$ must be an integer. Also notice that (NC) fails whenever n is a multiple of k.

2. Results. In [C] Chung, Diaconis, and Graham made the following conjecture, for which $100 is offered.

CONJECTURE 1. *For all k there is an integer $n_0(k)$ such that, for $n \geq n_0(k)$, Ucycles for k-subsets of $[n]$ exist if and only if* (NC) *holds.*

In [J1] we find the following results.

THEOREM 2. *Ucycles for* 3*-subsets of* $[n]$ *exist for all $n \geq 8$ not divisible by* 3.

THEOREM 3. *Ucycles for* 4*-subsets of* $[n]$ *exist for all odd $n \geq 9$.*

Theorem 2 is a verification of conjecture 1 for $k = 3$.

In order to verify the conjecture for $k = 4$ one would need to construct Ucycles for $n \equiv 2 \bmod 8$. No examples of this case have ever been found, nor have any with $k = 5$. The purpose of this paper is to prove the following theorem.

THEOREM 4. *For $k = 3, 4, 6$ there is an integer $n_0(k)$ such that, for $n \geq n_0(k)$, Ucycles for k-subsets of $[n]$ exist whenever n is relatively prime to k.*

This theorem includes both results of Jackson and unifies them in one setting together with the case $k = 6$.

3. Terminology. Let $S = \{s_1, \ldots, s_k\}, s_i < s_{i+1}$, be a k-subset of $[n]$. Define its *ordered difference set,* or d-set, $D(s) = (d_1, \ldots, d_k)$ by $d_i = s_{i+1} - s_i$, where indices are modulo k and arithmetic is modulo n. Two d-sets are *equivalent* if one is a cyclic permutation of the other. Thus, we say that any additive translation of S belongs to the same d-set as S, where an *additive translation* of S is any set $S + r = \{s_1 + r, \ldots, s_k + r\}$ modulo n. Two d-sets will belong to the same *d-class* whenever one is any permutation of the other. Thus the family of all d-classes is simply the family of all unordered partitions of the integer n into k parts. Each d-class in turn defines a partition of k according to the number of parts of the same size. We say that two d-classes belong to the same *d-pattern* if they define the same partition of k. A d-pattern is *good* if some part has size 1 and *bad* otherwise. Let us offer some examples with $k = 5$ and $n = 40$.

There are 7 d-patterns $\langle 1,1,1,1,1\rangle, \langle 2,1,1,1\rangle, \langle 3,1,1\rangle, \langle 4,1\rangle, \langle 5\rangle, \langle 2,2,1\rangle$, and $\langle 3,2\rangle$, of which only $\langle 5\rangle$ and $\langle 3,2\rangle$ are bad. $\langle 3,2\rangle$ contains the d-classes $[2,2,2,17,17], [4,4,4,14,14], \ldots,$ and $[12,12,12,2,2]$. $[2,2,2,17,17]$ contains the d-sets $(2,2,2,17,17)$ and $(2,2,17,2,17)$. $(2,2,17,2,17)$ contains the sets $\{1,3,5,22,24\}$, $\{2,4,6,23,25\}, \ldots,$ and $\{40,2,4,21,23\}$. The notation of braces, parentheses, brackets, and angles will be maintained throughout to distinguish the objects from one another.

4. Proof of Theorem 4. We first prove the following lemma.

LEMMA 5. *No bad d-patterns exist if and only if $k = 3, 4,$ or* 6 *and* $\gcd(n, k) = 1$.

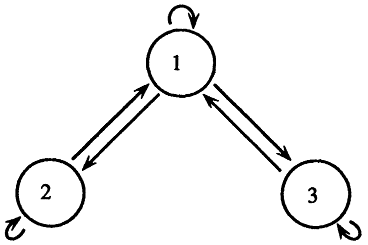


FIG. 1. $\mathcal{T}_{8,3}$.

Proof. If $r = \gcd(n,k)$ then the d-pattern $\langle r, r, \dots, r\rangle$ is bad unless $r = 1$. If $k > 3$ is odd then $\langle t, t+1\rangle$ is bad, where $k = 2t+1$ and $t > 1$, and if $k > 6$ is even then $\langle 3, t, t+1\rangle$ is bad, where $k = 2t+4$ and $t > 1$. Finally, it is easy to verify that if $k = 3, 4$, or 6 and $\gcd(n,k) = 1$ then no bad d-patterns exist. □

Next we define the *transition graph* $\mathcal{T}_{n,k}$, which is dependent upon the choices of representations for each d-class. Given the d-set $(d_1, \dots, d_{k-1}, d_k)$ we distinguish one of its coordinates (which we may assume to be d_k because of cyclic permutations) in the representation $(d_1, \dots, d_{k-1}; d_k)$ so as to imply the ordering $\{i, i+d_1, \dots, i+d_1+\dots+d_{k-1}\}$ of all its sets. The use of the semicolon in the representation is to identify which of the k differences is not to be used.

The representation of a d-class by $[d_1, \dots, d_{k-1}; d_k]$ signifies that d_k is distinguished (unused) in each of its d-sets. Thus, to avoid ambiguity, it is important that d_k be unique, in particular that the corresponding d-pattern be good. For example, with $k = 4$ and $n = 11$, we can choose $[2,2,1;6]$ to represent $[2,2,1,6]$. This determines the representations $(2,2,1;6), (2,1,2;6)$, and $(1,2,2;6)$ of its three d-sets, of which $(2,1,2;6)$ denotes the (ordered) sets $\{1,3,4,6\}, \{2,4,5,7\}, \dots$ and $\{11,2,3,5\}$.

Given the d-set representation $(d_1, \dots, d_{k-1}; d_k)$ we call the terms $((d_1, \dots, d_{k-2}))$ its *prefix* and $((d_2, \dots, d_{k-1}))$ its *suffix*. We use double parentheses to denote that these are vertices in the transition graph $\mathcal{T}_{n,k}$ whose directed edges are precisely the representations involved.

For example, Fig. 1 shows the transition graph $\mathcal{T}_{8,3}$ which was used to construct the Ucycle for 3-subsets of $[8]$ above. The d-sets are represented by $(1,1;6), (2,2;4), (3,3;2), (1,2;5), (2,1;5), (1,3;4)$, and $(3,1;4)$. The d-set $(2,1;5)$ corresponds to the directed edge $((2)) \to ((1))$, and so on. The Eulerian circuit 2211331 corresponds to a listing of all d-sets and produces the differences in the first block, 1356725, along with the first digit, 6, of the next block. Since the sum $2+2+1+1+3+3+1 \equiv 5 \bmod 8$, each block shifts by 5, and since 5 is relatively prime to 8 each integer occurs as the starting digit of some block. Hence, each 3-subset of $[8]$ occurs exactly once. As we shall see, it is an unnecessary luxury that the sum (5) is relatively prime to n (8).

LEMMA 6. *If $\mathcal{T}_{n,k}$ is Eulerian for some choice of representations of d-classes, then there exists a Ucycle for k-subsets of $[n]$.*

Proof of Lemma 6. $((1,1,\dots,1))$ is always a vertex of $\mathcal{T}_{n,k}$ and there is always a loop at that vertex representing the d-set $(1,1,\dots,1;d)$. This is because there is no other way to represent the d-class $[1,1,\dots,1,d]$, because we would be unable to distinguish which 1 would be set apart by a semicolon if we tried. Let $S_1 = \{1,2,\dots,k\}, S_2 = \{2,3,\dots,k+1\} = S_1 + 1$, and in general, let $S_{i+1} = S_i + 1$ be the sets belonging to this representation. Starting at S_i we follow along the edges of an Eulerian circuit in $\mathcal{T}_{n,k}$ until we return to our original loop. Repeating the loop now determines the set $S_i + r$ for some $0 \le r < n$. Retrace the circuit repeatedly until we finally reach S_i again, now having used the sets $S_i, S_i + r, S_i + 2r, \dots$, and $S_i - r$, in that order. Call the cycle produced U_i. Now let $s = \gcd(r, n)$. Then if $s = 1$ we have actually produced our Ucycle since we have used each of the n sets from every d-set (our initial example with $n = 8$ and $k = 3$ has $r = 5$ and $s = 1$). Otherwise, we have constructed s disjoint cycles $U_1, U_2, \dots, U_s$, which must somehow be hooked together to form one. We do this by a method we call *insertion,* showing how to insert U_{i+1} into U_i. Actually, the general case is the same as the first, so we merely show how to insert U_2 into U_1. Given

$$U_1 = 1, 2, \dots, k-1, k, x, \dots, y$$

and

$$U_2 = 2, 3, \dots, k, k+1, x+1, \dots, y+1$$

we insert as below.

$$1, \underline{2, 3, \dots, k, k+1, x+1, \dots, y+1}, 2, 3, \dots, k-1, k, x, \dots, y$$

Induction then completes the proof. □

Next we take a closer look at $\mathcal{T}_{n,k}$ by defining the graph $\mathcal{T}_{n,k}(C)$ for each d-class $C = [d_1, \dots, d_{k-1}; d_k]$. It is merely the restriction of $\mathcal{T}_{n,k}$ to edges that are representations of d-sets belonging to C. We now introduce the *class graph* $\mathcal{H}_{n,k}$, whose vertices are all possible d-classes and whose undirected edges join d-classes whose representations differ by one entry. For example, $[2, 2, 1, 3, 6; 7]$ and $[1, 1, 2, 2, 3; 12]$ are connected in $\mathcal{H}_{21,6}$.

LEMMA 7. *If $\mathcal{H}_{n,k}$ is connected and there are no bad d-patterns for k-subsets of $[n]$, then $\mathcal{T}_{n,k}$ is Eulerian.*

Proof of Lemma 7. If the d-class $C = [d_1, \dots, d_{k-1}; d_k]$ belongs to a good d-pattern, then $\mathcal{T}_{n,k}(C)$ is a disjoint union of cycles. This is so because no d_i equals d_k and so all distinct permutations of the entries $d_1, \dots, d_{k-1}$ can be partitioned into cyclic permutations of a fixed few. For example, with $C = [1, 1, 2, 2, 3; 12]$ we obtain the d-sets $(1, 1, 2, 2, 3; 12), (1, 1, 2, 3, 2; 12), (1, 1, 3, 2, 2; 12), (1, 2, 1, 2, 3; 12), (1, 2, 1, 3, 2; 12)$, and $(1, 3, 1, 2, 2; 12)$, along with all their five cyclic permutations each. Those of the last d-set induce the cycle of edges

$$\begin{aligned} ((1,3,1,2,2)) &\to ((3,1,2,2,1)) \to ((1,2,2,1,3)) \\ &\to ((2,2,1,3,1)) \to ((2,1,3,1,2)) \to ((1,3,1,2,2)) \end{aligned}$$

in $\mathcal{T}_{21,6}(C)$. An example of why this breaks down for bad d-patterns is the d-class $C = [1, 1, 4, 4]$ for 4-subsets of $[10]$, containing only the two d-sets $(1, 1, 4, 4)$ and $(1, 4, 1, 4)$. If we try to use the representation $[1, 1, 4; 4]$, then that induces the representations $(1, 1, 4; 4)$ and $(1, 4, 1; 4)$, which are the edges $((1, 1)) \to ((1, 4)) \to ((4, 1))$ in $\mathcal{T}_{10,4}(C)$. We can never form cycles this way.

We quickly see from this that if $\mathcal{T}_{n,k}$ is connected, then it is Eulerian, being a union of cycles. Two d-classes C_1 and C_2 being connected in $\mathcal{H}_{n,k}$ means that $\mathcal{T}_{n,k}(C_1)$ and $\mathcal{T}_{n,k}(C_2)$ share many of the same vertices. In particular, if C_1 and C_2 share $d_1, \dots, d_{k-2}$, then their corresponding graphs share every permutation of them as vertices. For example, $((2, 2, 1, 3)), ((2, 2, 3, 1))$, and $((2, 1, 2, 3))$ are the shared vertices of $[2, 2, 1, 3, 6; 7]$ and $[1, 1, 2, 2, 3; 12]$ above. However, this does not mean that the union of the two graphs $\mathcal{T}_{n,k}(C_1)$ and $\mathcal{T}_{n,k}(C_2)$ is connected, although since $\mathcal{H}_{n,k}$ is connected, we do get paths to and from every d-set and $(1, 1, \dots, 1)$, so that the union over all d-classes produces the connected graph $\mathcal{T}_{n,k}$. □

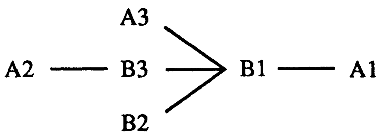


FIG. 2. *Connectedness of* $\mathcal{H}_{n,3}$.

LEMMA 8. *Let* $n_0(3) = 8, n_0(4) = 9,$ *and* $n_0(6) = 17$. *Then* $\mathcal{H}_{n,k}$ *is connected for* $k = 3, 4$ *and* 6 *with* $n \geq n_0(k)$ *and* $\gcd(n, k) = 1$.

Proof of Lemma 8.

Case k = 3. We assume $n \geq 8$ and break up the d-classes into the following types.

(A) d-pattern $\langle 2, 1\rangle : [a, a; b]$
(1) $a = 1$
(2) $a > 1, b = 1$ (for n odd only)
(3) $a > 1, b > 1$
(B) d-pattern $\langle 1, 1, 1\rangle : [a, b; c]$
(1) $1 = a < b < c$
(2) $1 < a < b < c$
(3) $2 = a < b - 1 = c$ (for n odd only)

We show that $\mathcal{H}_{n,3}$ is connected by displaying the path of edges from the d-classes of each type to the type A1 in Fig. 2. If n is odd, then A2 → B3 → B1 by $[a, a; 1] \to [2, a; a-1] \to [1, 2; c]$. Then A3 → B1 by $[a, a; b] \to [1, a; c]$, B2 → B1 by $[a, b; c] \to [1, b; d]$, and B1 → A1 by $[1, b; c] \to [1, 1; d]$.

Case k = 4. We assume $n \geq 9$ and break up the d-classes into the following types.

(A) d-pattern $\langle 3, 1\rangle : [a, a, a; b]$
(1) $a = 1$
(2) $a > 3, b = 1$
(3) $a > 1, b > 1$
(B) d-pattern $\langle 2, 1, 1\rangle : [a, a, b; c]$
(1) $a = 1, 1 < b < c$
(2) $a > 1, 1 = b < c$
(3) $a > 1, 1 < b < c$
(C) d-pattern $\langle 1, 1, 1, 1\rangle : [a, b, c; d]$
(1) $1 = a < b < c < d$
(2) $1 < a < b < c < d$

We show $\mathcal{H}_{n,4}$ is connected by displaying the path of edges from the d-classes of each type to the type A1 in Fig. 3. A2 → B3 → C1 by $[a, a, a; 1] \to [a, a, 2; a-1] \to [1, 2, a; b]$, C2 → C1 → B1 by $[a, b, c; d] \to [1, b, c; e] \to [1, 1, c; f]$, A3 → B2 → B1 by $[a, a, a; b] \to [a, a, 1; c] \to [1, 1, a; d]$, and B1 → A1 by $[1, 1, a; b] \to [1, 1, 1; c]$.

Case $k = 6$.

We assume $n \geq 17$ and break up the d-classes into the following types.

(A) d-pattern $\langle 5, 1\rangle : [a, a, a, a, a; b]$
(1) $a = 1$
(2) $a > 1, b > 1$
(3) $a > 1, b = 1$
(B) d-pattern $\langle 4, 1, 1\rangle : [a, a, a, a, b; c]$

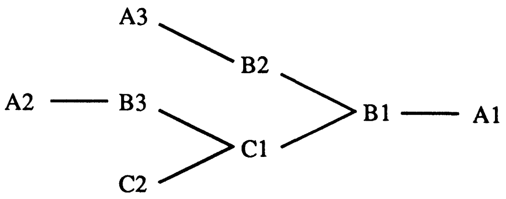


FIG. 3. *Connectedness of* $\mathcal{H}_{n,4}$.

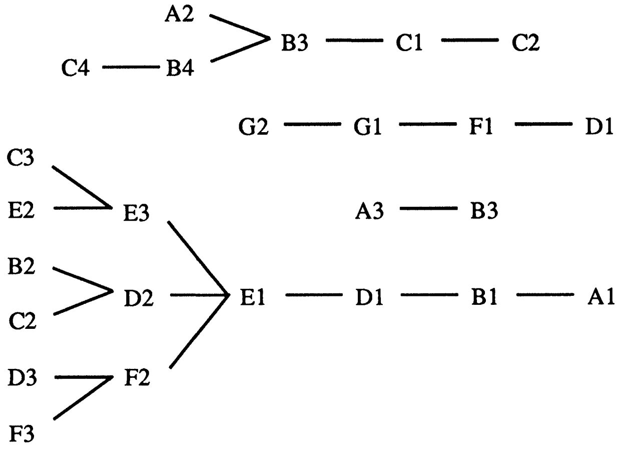


FIG. 4. *Connectedness of* $\mathcal{H}_{n,6}$.

(1) $a = 1, 1 < b < c$
(2) $a > 1, 1 < b < c - 2$
(3) $a > 1, 1 = b < c - 2$
(4) $a > 1, b = c + 1$

(C) d-pattern $\langle 3, 2, 1, 1\rangle : [a, a, a, b, b; c]$
(1) $a > b$
(2) $a < b, c > 1$
(3) $2 < a < b, c = 1$
(4) $2 = a < b, c = 1$

(D) d-pattern $\langle 3, 1, 1, 1\rangle : [a, a, a, b, c; d]$
(1) $a = 1, 1 < b < c < d$
(2) $a > 1, 1 = b < c < d$
(3) $a > 1, 1 < b < c < d$

(E) d-pattern $\langle 2, 2, 1, 1\rangle : [a, a, b, b, c; d]$
(1) $1 = a < b, 1 < c < d$
(2) $1 < a < b, 1 < c < d$
(3) $1 < a < b, 1 = c < d$

(F) d-pattern $\langle 2, 1, 1, 1, 1\rangle : [a, a, b, c, d; e]$
(1) $a = 1, 1 < b < c < d < e$
(2) $a > 1, 1 = b < c < d < e$
(3) $a > 1, 1 < b < c < d < e$

(G) d-pattern $\langle 1, 1, 1, 1, 1, 1\rangle : [a, b, c, d, e; f]$
(1) $1 = a < b < c < d < e < f$
(2) $1 < a < b < c < d < e < f$

We show $\mathcal{H}_{n,6}$ is connected by displaying the path of edges from the d-classes of each type to the type A1 in Fig. 4. C4 → B4 → B3 by $[2, 2, 2, b, b; 1] \to [2, 2, 2, 2, b; b - 1] \to [2, 2, 2, 2, 1; c]$, A2 → B3 → C1 by $[a, a, a, a, a; b] \to [a, a, a, a, 1; c] \to [a, a, a, 1,$

$1;d]$, A3 → B2 by $[a,a,a,a,a;1] \to [a,a,a,a,2;a-1]$, and C1 → C2 → D2 by $[a,a,a,b,b;c] \to [b,b,b,a,a;d] \to [b,b,b,1,a;e]$. B2 → D2 → E1 by $[a,a,a,a,b;c] \to [a,a,a,1,b;d] \to [1,1,a,a,b;e]$, C3 → E3 by $[a,a,a,b,b;1] \to [a,a,b,b,1;c]$, and E2 → E3 → E1 by $[a,a,b,b,c;d] \to [a,a,b,b,1;e] \to [1,1,a,a,b;f]$. D3 → F2 by $[a,a,a,b,c;d] \to [a,a,1,b,d;e]$, F3 → F2 → E1 by $[a,a,b,c,d;e] \to [a,a,1,c,d;f] \to [1,1,a,a,d;g]$, and G2 → G1 → F1 by $[a,b,c,d,e;f] \to [1,b,c,d,e;g] \to [1,1,c,d,e;h]$. And finally E1 → D1 by $[1,1,b,b,c;d] \to [1,1,1,b,c;d]$, F1 → D1 by $[1,1,b,c,d;e] \to [1,1,1,c,d;f]$, and D1 → B1 → A1 by $[1,1,1,b,c;d] \to [1,1,1,1,b;e] \to [1,1,1,1,1;f]$. □

Proof of Theorem 4. The proof follows from Lemmas 5–8. □

We finally note that Jackson [J2] has recently found by computer Ucycles for the values of (n,k) equal to $(10,4), (8,5), (9,5), (10,6), (11,6), (10,7), (11,7)$, and $(11,8)$.

REFERENCES

[B] N. G. de Bruijn, *A combinatorial problem*, Nederl. Akad. Wetensch. Proc., 49 (1946), pp. 758–764.

[C] F. R. K. Chung, P. Diaconis, and R. I. Graham, *Universal cycles for combinatorial structures*, Discrete Math., 110 (1992), pp. 43–59.

[F] H. Fredericksen, *A survey of full length nonlinear shift register cycle algorithms*, SIAM Rev., 24 (1982), pp. 195–221.

[G] I. J. Good, *Normally recurring decimals*, J. London Math. Soc., 21 (1946), pp. 167–169.

[H] G. Hurlbert, *Universal Cycles: On beyond De Bruijn*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1990.

[J1] B. Jackson, *Universal cycles for k-subsets and k-permutations*, Discrete Math., 117 (1993), pp. 141–150.

[J2] B. Jackson, personal communication May 1993.

[M] C. Flye-Sainte Marie, *Solution to problem number* 58, l'Intermediaire des Mathematicieus, 1 (1894), pp. 107–110.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 605–613, November 1994

009

FURTHER RESULTS ON T-COLORING AND FREQUENCY ASSIGNMENT PROBLEMS *

ARUNDHATI RAYCHAUDHURI†

Abstract. A graph coloring problem (called T-coloring) is investigated in which integers are assigned to the vertices of a graph G, under the constraint that the absolute value of the difference between integers assigned to adjacent vertices does not belong to a forbidden set (called the T-set). The T-coloring problem has applications in the radio frequency channel assignment problem. T-sets of the form $\{0, s, 2s, \ldots, ks\} \cup S$, where $s, k \geq 1$ and $S \subseteq \{s+1, s+2, \ldots, ks-1\}$, are considered. This set has been named the k-multiple of s set. Values of interest are the minimum cardinality, $\chi_T(G)$, and the minimum span, $\mathrm{sp}_T(G)$, of the set of assigned integers, over all possible T-colorings of G. A greedy algorithm, which T-colors perfectly orderable graphs in $\chi_T(G)$ colors with span $\mathrm{sp}_T(G)$ (for a k-multiple of s set), is given.

Key words. vertex coloring, T-coloring, frequency assignment, minimum span, greedy algorithm, perfectly orderable graphs

AMS subject classifications. 05C15

1. Introduction and definitions.

Efficient frequency assignment, which uses a minimum number or a minimum span of channels or frequencies assigned to a system of transmitters, has been motivated by the increasing scarcity of available radio frequency spectrum. In this context, T-coloring was first introduced by Hale [1980], who formulated in graph-theoretic terms the most general form of the channel assignment problem, with k levels of interference. In this paper we investigate this T-coloring problem when assigning channels for one level of interference only. We describe the problem below.

There are n transmitters $x_1, x_2, \ldots, x_n$ situated in a region. To each transmitter x_i, a channel $f(x_i)$ (a fixed positive integer), is to be assigned. Some of the transmitters interfere because of proximity, meteorological, or other reasons. Two interfering transmitters must be given frequencies such that the absolute value of the difference of their frequencies does not belong to a forbidden set T of nonnegative integers. Our objective is to make a frequency assignment that is efficient according to certain criteria (to be defined later in this section), while satisfying the above constraint.

To formulate this problem graph-theoretically, we define a graph G with $V(G) = \{x_1, x_2, \ldots, x_n\}$, and an edge between transmitters x_i and x_j iff they interfere. Given G and a T-set, a set of nonnegative integers, a T-coloring for G is a function f: $V(G) \rightarrow Z^+$ (the set of positive integers), such that

$$\{x, y\} \in E(G) \rightarrow |f(x) - f(y)| \notin T.$$

We assume that $0 \in T$, otherwise the problem becomes trivial. However, if $T = \{0\}$, then the problem reduces to the ordinary coloring problem.

* Received by the editors October 2, 1989; accepted for publication (in revised form) December 27, 1993.

† Department of Mathematics, College of Staten Island (CUNY), 2800 Victory Boulevard, Staten Island, New York 10314. Portions of this work are included in the author's Ph.D. dissertation completed under the supervision of Prof. Fred S. Roberts, Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

Previous work on the topic of T-coloring has been done by Cozzens and Roberts [1982]. They have studied T-sets of the form $\{0, 1, 2, \ldots, r\} \cup S$, where S is any set that does not contain any multiple of $(r+1)$. They call this set an *r-initial set.* Cozzens and Wang [1984] investigated T-sets where S may include some multiples of $(r+1)$. Tesman [1989] studied the T-coloring problem with two interference levels on nested unit interval graphs with corresponding T-sets $\{0\}$ and $\{0,1\}$. More recently D. Liu has given a useful characterization of T-sets for which $\mathrm{sp}_T(G) = \mathrm{sp}_T(K_{\chi(G)})(\mathrm{sp}_T(G) \leq \mathrm{sp}_T(K_{\chi(G)}))$ for any graph G (see Thm. 1(a), (ii)). A useful survey on generalized graph coloring including T-coloring and their applications can be found in Roberts [1989].

In the case of radio frequency assignment the forbidden T-sets can be very complex and difficult to model (see Middlekamp [1978] for a discussion of such T-sets). Here we shall primarily deal with one such T-set, the k-multiple of s set. The *k-multiple of s set* is of the form $\{0, s, 2s, \ldots, ks\} \cup S$, where s and $k \geq 1$ and $s \subseteq \{s+1, s+2, \ldots, ks-1\}$. Some practical forbidden sets, such as those that arise in UHF-TV, are very similar to k-multiple of s sets.

For any T-set, it is easy to find a T-coloring. If α is the largest entry of T, then one could use a different channel for each transmitter, choosing channels from $\{1, \alpha+2, 2\alpha+3, \ldots\}$. However this will not be very efficient in terms of the difference between the smallest and largest channels used. Our objective is to find efficient assignments for various graphs and T-sets. So let us at this point define some criteria for efficient channel assignment.

Given any graph G and a T-set, the *order* of a T-coloring f is the number of distinct integers $f(x)$, and the *span* is the maximum value of $|f(x) - f(y)|$ over all vertices x and y. The *minimum order* (or the *T-chromatic number*) and the *minimum span* (or the *T-span*) over all T-colorings for any graph G and a given T-set T are denoted by $\chi_T(G)$ and $\mathrm{sp}_T(G)$, respectively.

It is interesting to note that the above two criteria of optimization may not be satisfied by a single T-coloring, as pointed out by Cozzens and Roberts [1982]. The example in Fig. 1 illustrates this. Here G is a cycle on five vertices and $T = \{0, 1, 2, 6\}$. Figure 1(a) shows a minimum span assignment (of span 6) that uses five colors, the least possible number of colors to attain this span, and Fig. 1(b) shows a minimum order assignment that uses three colors and has a span of 7, the least possible span to attain this order.

This motivates us to define the restricted span and the restricted chromatic number of a T-coloring. *Restricted span* or $\mathrm{rsp}_T(G)$ is the minimum span of a T-coloring of G which uses $\chi_T(G)$ colors. *Restricted chromatic number,* $\chi_T^r(G)$, is the minimum order of a T-coloring of G whose span is $\mathrm{sp}_T(G)$.

The rest of this paper is organized as follows. Section 2 presents a study of $\mathrm{sp}_T(K_n)$ and $\mathrm{sp}_T(G)$, where T is a k-multiple of s set. In §3 we discuss the numbers $\chi_T^r(G)$ and $\mathrm{rsp}_T(G)$. In §4 we give a greedy algorithm that colors perfectly orderable graphs in $\chi_T(G)$ colors for any T-set, with span $\mathrm{sp}_T(G)$, if T is an r-initial set or a k-multiple of s set.

Additional notation used in this work are defined below:

$\chi(G)$	The chromatic number of G
K_n	A complete graph on n vertices
$\omega(G)$	Maximum size of a clique in G.

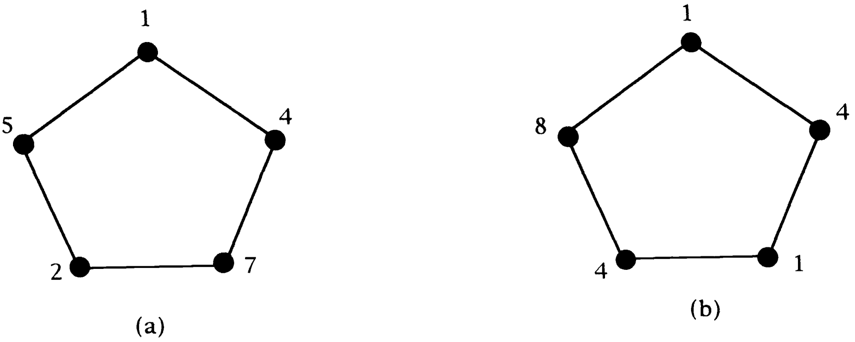


FIG. 1. *Two T-colorings of G for $T = \{0,\ 1,\ 2,\ 6\}$ for two different optimality criteria.*

We end this section by mentioning a result of Cozzens and Roberts [1982] involving $\chi_T(G)$ and $\mathrm{sp}_T(G)$, which we have used extensively in this paper.

THEOREM 1 (a) (Cozzens and Roberts [1982]). *For all graphs G and all T-sets T,*

(i) $\chi_T(G) = \chi(G)$,

and if $\chi_T(G) = m$ *and* $\omega(G) = l$, *then*

(ii) $\mathrm{sp}_T(K_l) \leq \mathrm{sp}_T(G) \leq \mathrm{sp}_T(K_m)$.

THEOREM 1 (b) (Cozzens and Roberts [1982]). *If T is an r-initial set, then* $\mathrm{sp}_T(G) = \mathrm{sp}_T(K_{\chi(G)}) = (r+1)(\chi(G)-1)$.

2. The numbers $\mathrm{sp}_T(K_n)$ and $\mathrm{sp}_T(G)$, where T is a k-multiple of s set.

In this section we shall study the numbers $\mathrm{sp}_T(K_n)$ and $\mathrm{sp}_T(G)$ where G is any graph and the T-set is a k-multiple of s set. Since by Theorem 1(a), (ii), $\mathrm{sp}_T(G)$ lies between the minimum spans of two complete graphs we shall first concentrate on finding $\mathrm{sp}_T(K_n)$ for any complete graph K_n. In Theorem 2 we shall show that for a k-multiple of s T-set we can produce a T-coloring of K_n of span $\mathrm{sp}_T(K_n)$ if we use a greedy algorithm. In Theorem 3 we shall show that $\mathrm{sp}_T(G)$ equals $\mathrm{sp}_T(K_{\chi(G)})$ (the upper bound for $\mathrm{sp}_T(G)$ as stated in Theorem 1(a), (ii)), where T is a k-multiple of s set. So we can conclude that for any G with known chromatic number $\chi(G)$, and a k-multiple of s T-set, we can produce a T-coloring of G of span $\mathrm{sp}_T(G)$ using a greedy algorithm.

The greedy algorithm for T-coloring is as follows. Let $x_1, x_2, \ldots, x_v$ be any vertex ordering of G. Color x_1 with 1, i.e., let $f(x_1) = 1$. Having assigned colors $f(x_1), f(x_2), \ldots, f(x_p)$, let $f(x_{p+1})$ be the smallest positive integer such that $|(x_{p+1}) - f(x_k)| \notin T$, where $k \leq p$ and $\{x_{p+1}, x_k\} \in E(G)$. Note that this algorithm produces a legitimate T-coloring of any graph G and any T-set T. However, it may not always give a coloring of minimum order or minimum span. For example, if $T = \{0,1,4,5\}$, and G is a K_3, the greedy algorithm uses colors $1,3,9$ whereas colors $1,4,7$ give a minimum span T-coloring. The following theorem is the main result of this section, which says that in case of a k-multiple of s T-set the greedy algorithm does produce a coloring of K_n of minimum span.

THEOREM 2. *Suppose* $T = \{0,\ s,\ 2s,\ 3s, \ldots,\ ks\} \cup S$, *where* $s,\ k \geq 1$, *and* $S \subseteq \{s+1, \ldots,\ ks-1\}$. *Then the greedy algorithm produces a T-coloring of* K_n *of span* $=$ $\mathrm{sp}_T(K_n)$. *Moreover, if* $n = st$, $t \in \{1,\ 2,\ 3, \ldots\}$, *then* $\mathrm{sp}_T(K_n) = st + skt - sk - 1$, *and if* $n = st + l$, $l \in \{1,\ 2, \ldots,\ s-1\}$, *then* $\mathrm{sp}_T(K_n) = st + skt + l - 1$.

Proof. Let $G = K_n$ and T be a k-multiple of s set. Color the vertices of G with the following colors, until all the vertices have been colored.

$$\begin{aligned}&1, 2, 3, 4, \ldots, s-1, s,\\ &(k+1)\,s+1, (k+1)\,s+2, \ldots, (k+2)\,s,\\ &(k+1)\,2s+1, (k+1)\,2s+2, \ldots, (2k+3)\,s, \text{etc.}\end{aligned}$$

That is, start with colors $1, 2, \ldots, s$, then skip ks colors to $s + ks + 1 = (k + 1)s + 1$ and use consecutive colors through $(k+1)s+s = (k+2)s$, then skip another ks colors to $(k + 1)2s + 1$ and so on. For $j = 1, 2, \ldots$, the $(js + 1)$th color has a difference of $ks + 1$ from the jsth color. This is a valid T-coloring because the difference between any two colors from the above set $\in \{1, 2, \ldots, s - 1\}$, or is a number $\geq ks + 1$. Note that in choosing these colors for T-coloring K_n, we have used the greedy algorithm. We shall now find the span resulting from this T-coloring.

Suppose $n = st + l$; where $t \in \{1, 2, 3, \ldots\}$ and $l \in \{0, 1, 2, \ldots, s - 1\}$. Then the span of K_n as generated by the greedy algorithm can easily be seen as follows.

Case 1: $n = st$. Span of K_n (as generated by the greedy algorithm) $= [(t-1)k + t]s - 1 = st + skt - sk - 1$.

Case 2: $n = st + l : l \in \{1, 2, \ldots, s - 1\}$. Span of K_n (as generated by the greedy algorithm) $= [(t-1)k + t]s + ks + l - 1 = st + skt + l - 1$.

Next, we shall show that the above assignment is a minimum span assignment. Consider a valid T-coloring of K_n of span $= \mathrm{sp}_T(K_n)$. In this T-coloring two colors c_1 and c_2 are said to be of the same type if $c_1 \equiv c_2$ modulo s. Therefore if $c_1 = i_1 s + j$ and $c_2 = i_2 s + j$ are two colors of the same type (namely type j) in a valid T-coloring of K_n then $|c_1 - c_2| \geq s(k + 1)$, since the T-set is a k-multiple of s set.

If $n = st$ then there are st distinct colors in this coloring, since $0 \in T$. Also there are at most t colors of any type used. Otherwise, if there are $t + 1$ or more distinct colors of the same type, then $\mathrm{sp}_T(K_n) \geq ts(k+1)$, since any two colors of the same type must differ by at least $s(k + 1)$. However, $\mathrm{sp}_T(K_n) \leq st + skt - sk - 1 < st + skt - 2$. So no more than t colors of any type have been used. Since each type of color contains $\leq t$ colors and there are at most s types, it must be that out of the st distinct colors, there are exactly t colors of each type $1, 2, \ldots, s - 1, 0$.

If $n = st + l, l \in \{1, 2, \ldots, s - 1\}$, then we claim that at most $t + 1$ colors of any type have been used, since any color type with $\geq t + 2$ colors in it would imply that $\mathrm{sp}_T(K_n) \geq (t + 1)s(k + 1) = skt + st + sk + s \geq st + skt + s + 1$. But we know that $\mathrm{sp}_T(K_n) \leq st + skt + l - 1 \leq st + skt + s - 2$. Again, we can argue that since no type has $> t + 1$ colors and there are at most s types of colors, among the distinct $st + l$ colors there must be at least l types of colors $i_1, i_2, \ldots, i_l$ with exactly $t + 1$ colors of each type.

Let α_i be the smallest color of type $i, i \in \{1, 2, \ldots, s - 1, 0\}$ if $n = st$, and $i \in \{i_1, i_2, \ldots, i_l\}$ if $n = st+l$. Without loss of generality, $\alpha_1 < \alpha_2 < \cdots < \alpha_{s-1} < \alpha_0$, if $n = st$, and $\alpha_{i1} < \alpha_{i2} < \cdots < \alpha_{il}$ if $n = st + l$. Let β_0 and βi_l be the highest colors of types 0 and i_l, respectively, used in these two cases. Then, if $n = st, \mathrm{sp}_T(K_n) \geq (\beta_0 - \alpha_1) = (\beta_0 - \alpha_0) + (\alpha_0 - \alpha_1) \geq (t-1)s(k+1) + (s-1) = st + skt - sk - 1$. ($\beta_0, \alpha_0$ are two colors of the same type and there are exactly t colors of this type; also, $\alpha_0 - \alpha_1 \geq s - 1$ since $\alpha_1, \alpha_2, \ldots, \alpha_{s-1}, \alpha_0$ are distinct.) On the other hand, if $n = st + l$, then $\mathrm{sp}_T(K_n) \geq (\beta_{il} - \alpha_{il}) = (\beta_{il} - \alpha_{il}) + (\alpha_{il} - \alpha_{il}) \geq ts(k+1) + l - 1 = st + skt + l - 1$. ($\beta_{il}$ and α_{il} are two colors of the same type and there are exactly $t + 1$

colors of this type.) Therefore in both cases $n = st$ and $n = st+l, l \in \{1, 2, \ldots, s-1\}$, the spans of K_n produced by the greedy algorithm are as stated in Theorem 2. □

COROLLARY 2.1. *Suppose G is any graph and T is a k-multiple of s set. If $\omega(G) = st + m$ and $\chi(G) = sq + n, t, q \in \{1, 2, 3, \ldots\}$ and $m, n \in \{0, 1, \ldots, s-1\}$, then,*

(1) *if $m = 0$ and $n = 0$, then $st + skt - sk - 1 \le \mathrm{sp}_T(G) \le sq + skq - sk - 1$.*
(2) *if $m \neq 0$ and $n = 0$, then $st + skt + m - 1 \le \mathrm{sp}_T(G) \le sq + skq - sk - 1$.*
(3) *if $m = 0$ and $n \neq 0$, then $st + skt - sk - 1 \le \mathrm{sp}_T(G) \le sq + skq + n - 1$.*
(4) *if $m \neq 0$ and $n \neq 0$, then $st + skt + m - 1 \le \mathrm{sp}_T(G) \le sq + skq + n - 1$.*

Proof. By Theorem 1(a), (ii), $\mathrm{sp}_T(K_{\omega(G)}) \le \mathrm{sp}_T(G) \le \mathrm{sp}_T(K_{\chi(G)})$. □

THEOREM 3. *Suppose T is a k-multiple of s set. Then for any graph G, $\mathrm{sp}_T(G) = \mathrm{sp}_T(K_{\chi(G)})$.*

Proof. We shall prove this by showing that if f is any T-coloring of G of span $= \mathrm{sp}_T(G)$, we can change it to another T-coloring g of G, where $g(x) \le f(x)$ for all x in $V(G)$, minimum $g(x) =$ minimum $f(x)$, and $g(x)$ uses exactly those colors used by the greedy algorithm when coloring $K_{\chi(G)}$. Thus $\mathrm{sp}_T(G) =$ span of the T-coloring $g = \mathrm{sp}_T(K_{\chi(G)})$ by Theorem 2.

Let f be a T-coloring of G of span $= \mathrm{sp}_T(G)$. Without loss of generality f uses color 1. In order to define the new g T-coloring from the fT-coloring we shall partition the set of positive integers into certain sets as follows.

$I_{1,l}$ = The set of integers in the interval $[ls + (l-1)ks + 1, ls + lks]$ for $l = 1, 2, 3, \ldots$

$I_{2,l} = \{ls + lks + 1; l = 0, 1, 2, \ldots\}$

$I_{3,l} = \{ls + lks + 2; l = 0, 1, 2, \ldots\}$

. . . .

$I_{s,l} = \{ls + lks + (s-1); l = 0, 1, 2, \ldots\}$

$I_{0,l} = \{ls + (l-1)ks; l = 1, 2, 3, \ldots\}$.

Note that $I_{1,l} \cup I_{2,l} \cup \cdots \cup I_{0,l}$ cover the entire set of positive integers. This is because

$$\{I_{2,0} \cup I_{3,0} \cup \cdots \cup I_{s,0}\} = \{1, 2, \ldots, s-1\}$$
$$\{I_{0,1} \cup I_{1,1} \cup I_{2,1} \cup \cdots \cup I_{s,1}\} = \{s, [s+1, s+ks], s+ks+1, \ldots, s+ks+s-1\}$$
$$\{I_{0,2} \cup I_{1,2} \cup I_{1,3} \cup \cdots \cup I_{s,2}\}$$
$$= \{2s+ks, [2s+ks+1, 2s+2ks], 2s+2ks+1, \ldots, 2s+2ks+s-1\}$$

and so on.

Thus every $f(x)$ belongs to $I_{j,l}$ for some j in $\{0, 1, 2, \ldots, s\}$ and some l in $\{0, 1, 2, \ldots\}$. From $f(x)$ define another T-coloring $g(x)$ as follows.

If $f(x) \notin I_{1,l}$ let $g(x) = f(x)$.

If $f(x) \in I_{1,l}$ and $f(x) \equiv t$ modulo s, then let $g(x) =$ largest integer $i < f(x), i \notin I_{1,l}$ and $i \equiv t$ modulo s. Such an i always exists since $\{I_{2,0} \cup \cdots \cup I_{s,0} \cup I_{0,1}\}$ contains $\{1, 2, \ldots, s\}$ and the least positive integer in $I_{1,l}$ is $s+1$. We claim that $g(x)$ is again a valid T-coloring. We now prove this claim.

Note that for all x in $V(G), g(x) \notin I_{1,l}$. Hence for all $x_1, x_2 \in V(G), g(x_1)$ and $g(x_2)$ are of one of the following forms.

(i) $g(x_1), g(x_2)$ are of the same class (i.e., belong to $I_{j,l}$ for same $j, j \neq 1$). Then $|g(x_1) - g(x_2)| = (l_1 - l_2)s(1+k)$, where l_1, l_2 are nonnegative integers such that $l_1 - l_2 \in \{0, 1, 2, \ldots\}$. So $|g(x_1) - g(x_2)|$ is either 0 or a positive integral multiple of $(k+1)s$.

(ii) $g(x_1), g(x_2)$ are of different classes, but neither of them belong to $I_{0,l}$. Then $|g(x_1) - g(x_2)| = (l_1 - l_2)s(k+1) + (n_1 - n_2)$ where l_1, l_2 are nonnegative integers such that $(l_1 - l_2) \in \{0, 1, 2, \ldots\}$ and $n_1, n_2 \in \{1, 2, \ldots, s-1\}, n_1 \neq n_2$. Note that $l_1 - l_2 = 0$ implies that $n_1 > n_2$. So $|g(x_1) - g(x_2)|$ either $\in \{1, 2, \ldots, s-2\}$ or *is* $> ks$.

(iii) $g(x_1)$ and $g(x_2)$ are of different classes and one of them belongs to $I_{0,l}$, namely $g(x_1) \in I_{0,l}$ and $g(x_1) > g(x_2)$. Then $|g(x_1) - g(x_2)| = (l_1 - l_2)s(k+1) - ks - n$, where l_1, l_2 are nonnegative integers such that $(l_1 - l_2) \in \{1, 2, \ldots\}$ and $n \in \{1, 2, \ldots, s-1\}$. So $|g(x_1) - g(x_2)|$ either $\in \{1, 2, \ldots, s-1\}$ or $> (k+1)s$.

(iv) $g(x_1), g(x_2)$ are of different classes and $g(x_1) \in I_{0,l}$ and $g(x_2) > g(x_1)$. Then $|g(x_1) - g(x_2)| = (l_1 - l_2)s(k+1) + ks + n$ where l_1, l_2 are nonnegative integers such that $(l_1 - l_2) \in \{0, 1, 2, \ldots\}$ and $n \in \{1, 2, \ldots, s-1\}$. So $|g(x_1) - g(x_2)| > ks$.

It can be verified that in all the above cases $|g(x_1) - g(x_2)| \notin$ interval $[s, ks]$. Thus it will suffice to show that $\{x, y\} \in E(G) \to |g(x) - g(y)| \neq 0$. Among all the above cases, $|g(x) - g(y)|$ could equal 0 only in case (i) when $g(x)$ and $g(y)$ belong to some $I_{j,l}$ for same $j (\neq 1)$ and same l. This will be the case only if $|f(x) - f(y)| = 0$ which implies that $\{x, y\} \notin E(G)$. This proves the claim that $g(x)$ is a valid T-coloring.

Now, minimum $g(x) =$ minimum $f(x) = 1$, and $g(x) \leq f(x)$, for all x in $V(G)$. Thus the span of the T-coloring $g \leq$ span of the T-coloring $f = \mathrm{sp}_T(G)$. So the span of the T-coloring $g = \mathrm{sp}_T(G)$. Let us list all the colors in $I_{0,l} \cup I_{2,l} \cup \cdots \cup I_{s,l}$ in ascending order. Let $1 = \alpha_1 < \alpha_2 < \cdots$ be those colors. Note that these are exactly the colors used by the greedy algorithm when T-coloring K_n (where T is a k-multiple of s set) as explained in the proof of Theorem 2. So, by Theorem 2, the greedy algorithm produces a coloring of $K_{\chi(G)}$ of span $= \mathrm{sp}_T(K_{\chi(G)})$ and uses the colors $\alpha_1, \alpha_2, \ldots, \alpha_{\chi(G)}$. The T-coloring g of G described above also uses colors from the set $I_{0,l} \cup I_{2,l} \cup \cdots \cup I_{s,l}$, namely colors $1 = \alpha_{i(1)} < \alpha_{i(2)} < \cdots < \alpha_{i(p)}$, where $i(p) \geq p \geq \chi(G)$. Thus the span of the T-coloring $g = \alpha_{i(p)} - 1 \geq \alpha_{\chi(G)} - 1 = \mathrm{sp}_T(K_{\chi(G)})$ (by Theorem 2). So, $\mathrm{sp}_T(G) =$ span of the T-coloring $g \geq \mathrm{sp}_T(K_{\chi(G)})$. But, $\mathrm{sp}_T(G) \leq \mathrm{sp}_T(K_{\chi(G)})$ by Theorem 1 (a), (ii). Therefore $\mathrm{sp}_T(G) = \mathrm{sp}_T(K_{\chi(G)})$. □

3. Restricted span and restricted chromatic number.

In this section we shall study in some detail the numbers $\chi_T^r(G)$ and $\mathrm{rsp}_T(G)$ which we introduced in §1. It follows easily from the definitions that $\chi_T^r(G) = \chi(G)$iff $\mathrm{rsp}_T(G) = \mathrm{sp}_T(G)$ ($\mathrm{rsp}_T(G) = \mathrm{sp}_T(G) \leftrightarrow$ there is a T-coloring that uses $\chi(G)$ colors and has span equal to $\mathrm{sp}_T(G) \leftrightarrow \chi_T^r(G) = \chi(G)$). We have mentioned earlier that for some graphs G and some forbidden sets T, we cannot achieve by a single coloring the minimum order and the minimum span that are possible for that G and T. In this section, however, we show that for weakly χ-perfect graphs (graphs for which $\chi(G) = \omega(G)$) with any T-set, and for any G with an r-initial or a k-multiple of s T-set, both these efficiency criteria can be achieved in the same coloring. We state that result in Theorem 4.

THEOREM 4. *For any G and any T*, $\mathrm{rsp}_T(G) = \mathrm{sp}_T(K_{\chi(G)})$.

Proof. (a) $\mathrm{rsp}_T(G) \leq \mathrm{sp}_T(K_{\chi(G)})$.

Consider a T-coloring of $K_{\chi(G)}$ of span $= \mathrm{sp}_T(K_{\chi(G)})$. Suppose it uses colors $\alpha_1, \alpha_2, \ldots, \alpha_{\chi(G)}$. Consider any T-coloring of G that uses $\chi_T(G) = \chi(G)$ colors, namely $\beta_1, \beta_2, \ldots, \beta_{\chi(G)}$. Find another legitimate T-coloring g of G by interchanging the colors β_i with α_i. Then $\mathrm{rsp}_T(G) \leq$ span of T-coloring $g = \mathrm{sp}_T(K_{\chi(G)})$.

(b) $\mathrm{rsp}_T(G) \geq \mathrm{sp}_T(K_{\chi(G)})$.

Consider a T-coloring of G that uses exactly $\chi(G)$ colors (namely $\alpha_1 < \alpha_2 < \cdots < \alpha_{\chi(G)}$) of span $\mathrm{rsp}_T(G)$. Let us divide the vertices of G into $\chi(G)$ color classes of colors $\alpha_1, \alpha_2, \ldots, \alpha_{\chi(G)}$. Between any pair of color classes α_i and α_j, there must be an edge joining a vertex in color class α_i to a vertex in color class α_j. Otherwise we could

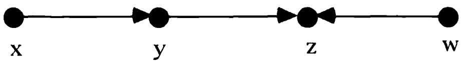


FIG. 2. *An obstruction.*

merge the two color classes resulting in a partition of $V(G)$ into $(\chi(G)-1)$ independent sets, which is a contradiction. So, $\alpha_1, \alpha_2, \ldots, \alpha_{\chi(G)}$ may form a T-coloring for $K_{\chi(G)}$. So, $\text{rsp}_T(G) = \alpha_{\chi(G)} - \alpha_1 \geq \text{sp}_T(K_{\chi(G)})$. □

COROLLARY 4.1. *For all weakly χ-perfect G and any T, $\text{rsp}_T(G) = \text{sp}_T(G)$ and $\chi_T^r(G) = \chi(G)$.*

Proof. By Theorem 4, $\text{rsp}_T(G) = \text{sp}_T(K_{\chi(G)})$ and by Theorem 1(a), (ii), for all weakly χ-perfect $G, \text{sp}_T(K_{\chi(G)}) = \text{sp}_T(G)$. So $\text{rsp}_T(G) = \text{sp}_T(G)$ and therefore $\chi_T^r(G) = \chi(G)$. □

COROLLARY 4.2. *For all G, and T an r-initial set or a k-multiple of s set, $\text{rsp}_T(G) = \text{sp}_T(G)$ and $\chi_T^r(G) = \chi(G)$.*

Proof. By Theorem 4, $\text{rsp}_T(G) = \text{sp}_T(K_{\chi(G)})$ and $\text{sp}_T(K_{\chi(G)}) = \text{sp}_T(G)$ for any G and T an r-initial set (by Theorem 1(b)) and for any G and T a k-multiple of s set (by Theorem 3). So for both these T-sets, $\text{rsp}_T(G) = \text{sp}_T(G)$ and consequently $\chi_T^r(G) = \chi(G)$. □

4. Greedy algorithms for perfectly orderable graphs. In this section we shall discuss T-coloring for perfectly orderable graphs, which is a very broad class of graphs including transitively orientable, rigid circuit, and complements of rigid circuit graphs. In Theorem 5 we shall show that for all perfectly orderable graphs G, a greedy algorithm results in a T-coloring of G in $\chi_T(G)$ colors. In particular if the T-set is either an r-initial or a k-multiple of s set then the span produced by this greedy algorithm equals $\text{sp}_T(G)$.

Let $x_1, x_2, \ldots, x_v$ be any ordering of $V(G)$. Using the greedy algorithm on this ordering, let us assign colors $f(x_j)$ to the vertices of G such that adjacent channels get different colors. Then the largest integer appearing as $f(xj)$ is called the *Grundy number* of this ordering. Obviously, the Grundy number of any ordering $\geq \chi(G)$. An *ordering* x_1, $x_2, \ldots, x_v$ *is perfect* if, for each generated subgraph H of G, the Grundy number of H is equal to $\chi(H)$. A graph G is *perfectly orderable* if $V(G)$ admits a perfect order.

An *ordering* x_1, $x_2, \ldots$, x_v on $V(G)$ is called *admissible* if the acyclic orientation of $E(G)$ implied by this ordering creates no generated obstruction, where an *obstruction* is the digraph shown in Fig. 2.

Chvatal [1981] showed that a graph G is perfectly orderable iff $V(G)$ admits an admissible order. He has also pointed out that transitively orientable graphs, chordal graphs (for definitions, see Golumbic [1980]), and complements of chordal graphs belong to the class of perfectly orderable graphs. In a transitively orientable graph, a transitive orientation creates no obstruction. In a chordal graph a reversed perfect elimination ordering is an admissible ordering. In a complement G^c of a chordal graph G, a perfect elimination ordering of $V(G)$ is an admissible ordering of $V(G^c)$.

LEMMA 1 (Chvatal [1981]). *Let G be a perfectly orderable graph with an admissible or perfect order $<$ on $V(G)$. Let Q be a clique in G such that each vertex w in Q has a neighbor $p(w)$ and these neighbors form an independent set. If $p(w) < w$ for all w in Q, then some $p(w)$ is adjacent to all vertices in Q.*

THEOREM 5. *Suppose G is a perfectly orderable graph and x_1, $x_2, \ldots$, x_v is a perfect ordering of $V(G)$. Then for all T-sets the greedy algorithm produces a T-coloring of G in $\chi_T(G) = \chi(G)$ colors.*

Proof. If we use the greedy algorithm to T-color the complete graph K_m, we shall use m different colors, say $1 = \alpha_1 < \alpha_2 < \cdots < \alpha_m$. Now if we apply the greedy algorithm to T-color the vertices of a perfectly orderable graph G (for which $\chi(G) = m$) using the order $x_1, x_2, \dots, x_v$, we shall show (by induction on $|V(G)| = v$) that we use at most the colors $\alpha_1, \alpha_2, \dots, \alpha_m$.

Clearly $f(x_1) = 1$. Suppose that the assertion is true for all perfectly orderable graphs G with $|V(G)| < v$. Take such a graph with $|V(G)| = v$ and T-color its vertices using the greedy algorithm. Consider the vertex x_v. Let $X_i^v = \{x \text{ in } \{x_1, x_2, \dots, x_{v-1}\}$: x_v is adjacent to x and x is of color $\alpha_i\}, i = 1, \dots, m$. If X_i^v is ϕ for any i, then the greedy algorithm will choose the smallest such α_i for x_v. Let us therefore assume that $X_i^v \neq \phi$ for any $i = 1, \dots, m$. Suppose that $z_i \in X_i^v$, for $i = 1, \dots, m$. Note that $\{z_1, z_2, \dots, z_m, x_v\}$ cannot form a clique since in that case $\omega(G) \geq m + 1 = \chi(G) + 1$. Therefore suppose that j is the smallest number among $1, 2, \dots, m$ such that $\{z_j, z_{j+1}, \dots, z_m, x_v\}$ forms a clique of G. By the preceding remark, $j > 1$. Let w be any vertex in $\{z_j, z_{j+1}, \dots, z_m, x_v\}$. Since the colors are assigned by the greedy algorithm, w must have a neighbor $p(w)$ in G which has been previously colored with color α_{j-1}. Since $p(w)$ was previously colored, the subscript of $p(w) <$ the subscript of w, i.e., $p(w)$ precedes w in the perfect ordering. Thus $\{z_j, z_{j+1}, \dots, z_m, x_v\}$ is a clique of G, each of whose members has a neighbor with a vertex labeled less than that of itself, and these neighbors form an independent set (since they are all colored with α_{j-1} and $0 \in T$). So by Lemma 1 one of these neighbors, say z_{j-1}, is adjacent to each element of $\{z_j, z_{j+1}, \dots, z_m, x_v\}$. Thus $\{z_{j-1}, z_j, z_{j+1}, \dots, z_m, x_v\}$ forms a clique, which is a contradiction to our assumption that the maximum clique size is m. Thus X_i^v must be empty for some $i = 1, 2, \dots, n$, so that one of the colors $\alpha_i, \alpha_2, \dots, \alpha_m$ will be available for coloring the vertex x_v.

Thus if we use the greedy algorithm we shall always T-color (for any T-set) a perfectly orderable graph G using colors from the set $\{\alpha_1, \alpha_2, \dots, \alpha_m\}$. Since $\chi_T(G) = \chi(G) = m$, the greedy algorithm uses exactly these colors and achieves a coloring with $\chi_T(G)$ colors. □

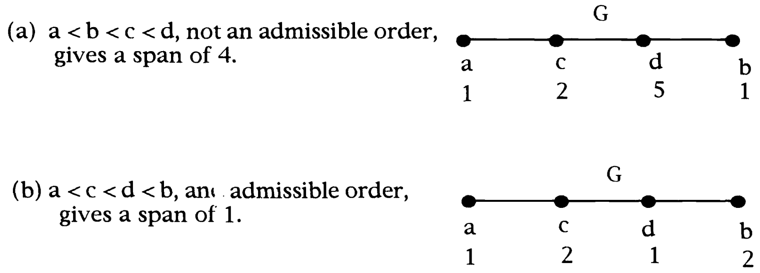
COROLLARY 5.1. *Suppose G is a perfectly orderable graph and $x_1, x_2, \dots, x_v$ is a perfect ordering of $V(G)$ and suppose T is either an r-initial set or a k-multiple of s set. Then the greedy algorithm produces a T-coloring of G of span equal to* $\mathrm{sp}_T(G)$.

Proof. (a) Suppose T is an r-initial set. By Theorem 5, the T-coloring of G produced by the greedy algorithm has span $= [(\chi(G)-1)r + \chi(G)] - 1 = (\chi(G)-1)(r+1)$ since it uses the same colors used for T-coloring $K_{\chi(G)}$ by the greedy algorithm. However, by Theorem 1(b), $\mathrm{sp}_T(G) = (\chi(G) - 1)(r + 1)$ for any G and T an r-initial set. Thus the span produced by this algorithm equals $\mathrm{sp}_T(G)$.

(b) Suppose T is a k-multiple of s set. Let us enumerate all the colors in the set $I_{0,l} \cup I_{2,l} \cup \cdots \cup I_{s,l}$ (as defined in the proof of Theorem 3) in ascending order. Let $1 = \alpha_1 < \alpha_2 < \cdots$ be such an ordering. These colors are used by the greedy algorithm to color any complete graph. So by Theorem 5, the T-coloring of G produced by the greedy algorithm has span $= \alpha_{\chi(G)} - 1$ since the colors used to T-color G are identical to those used for T-coloring $K_{\chi(G)}$ by the greedy algorithm. However, according to Theorem 3, $\mathrm{sp}_T(G) = \mathrm{sp}_T(K_{\chi(G)})$ and by Theorem 2, $\mathrm{sp}_T(K_{\chi(G)}) = \alpha_{\chi(G)} - 1$. Thus the span produced by this T-coloring $= sp_T(G)$. □

Remark. Note that Corollary 5.1 is applicable to chordal graphs, transitively orientable graphs, and complements of chordal graphs.

We illustrate Corollary 5.1 on Fig. 3 where we show a perfectly orderable graph G with an ordering $a < b < c < d$ of $V(G)$ which is not an admissible order. If $T = \{0, 2\}$, then the T-coloring of G produced by the greedy algorithm using this ordering has



(a) a < b < c < d, not an admissible order, gives a span of 4.

(b) a < c < d < b, an admissible order, gives a span of 1.

FIG. 3. *T-coloring of G by greedy algorithm for $T = \{0, 2\}$ using two orders.*

span equal to 4. However, the greedy algorithm on the ordering $a < c < d < b$, which is an admissible order, would give a T-coloring of span $= 1 = \text{sp}_T(G)$.

5. Concluding remarks. It has been shown that for any k-multiple of s T-set and any complete graph, the greedy algorithm leads to a T-coloring of minimum span with any ordering of the vertex set. The same is true for a perfectly orderable graph with a perfect ordering of the vertex set. For any other G and a k-multiple of s T-set, as long as $\chi(G)$ is known, we can find a T-coloring of minimum span. A constructive procedure is given for such a T-coloring in the proof of Theorem 3.

Acknowledgments. The author thanks the reviewers for their very helpful comments and suggestions. The author also thanks her thesis advisor, Professor F. S. Roberts, for his guidance and The College of Staten Island for partial research support.

REFERENCES

V. CHVATAL, *Perfectly ordered graphs*, McGill Univ. Tech. report SOCS-81, 28 (1981).

M. B. COZZENS AND F. S. ROBERTS, *T-colorings of graphs and the channel assignment problem*, Congr. Numer., 35 (1982), pp. 191–208.

M. B. COZZENS AND DING-I. WANG, *The general channel assignment problem*, Congr. Numer., 4 (1984), pp. 115–129.

M. C. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

W. K. HALE, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.

D. LIU, *T-colorings of graphs*, Discrete Math., 101 (1992), pp. 203–212.

L. C. MIDDLEKAMP, *UHF taboos—History and development*, IEEE Trans. Consumer Electronics, CE-24 (1978), pp. 514–519.

F. S. ROBERTS, *From garbage to rainbows: Generalization of graph coloring and their applications*, in Proc. Sixth International Conference on the Theory and Applications of Graphs, Y. Alavi, G. Chartrand, O. R. Ollermann, and A. J. Schwenk, eds., John Wiley, New York, 1989.

B. TESMAN, *T-colorings, list T-colorings, and set T-colorings of graphs*, Ph.D. thesis, Department of Mathematics, Rutgers University, New Brunswick, NJ, 1989.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 614–631, November 1994

010

A LINEAR-TIME ALGORITHM FOR ISOMORPHISM OF GRAPHS OF BOUNDED AVERAGE GENUS[*]

JIANER CHEN[†]

Abstract. A structure theorem is proved for the class of graphs of bounded average genus, which leads to a linear-time algorithm for isomorphism of such graphs.

Key words. algorithm, graph isomorphism, graph imbedding

AMS subject classifications. 05C10, 05C60, 05C85, 57M15, 68R10

1. Introduction. The problem of deciding whether two given finite graphs are isomorphic is known as the *graph isomorphism problem.* At present there is no known algorithm for determining if two arbitrary graphs are isomorphic with a running time that is asymptotically less than exponential. Finding the complexity of this problem remains an open problem on the list of Garey and Johnson [8].

It is known that for a few restricted classes of graphs, the isomorphism problem can be solved in polynomial time. Hopcroft and Tarjan [14] gave a linear-time algorithm for tree isomorphism and an $O(n \log n)$ time algorithm for planar graph isomorphism. A linear-time algorithm for isomorphism of planar graphs was later obtained by Hopcroft and Wong [15]. Filotti and Mayer [6] demonstrated that, for every genus graph, there is a polynomial-time algorithm whose degree rises with the genus. Furst, Hopcroft, and Luks [7] gave a subexponential-time algorithm for 3-regular graph isomorphism. Luks [16] improved this to show that for any fixed value of maximum valence, there is a polynomial-time algorithm.

A topological approach has recently been suggested by Gross and Furst [9] to solve the graph isomorphism problem with a probabilistic algorithm. They define a hierarchy of increasingly large topological invariants of isomorphism or homeomorphism type from which one might extract a polynomial-sized sample. Average genus is at the low end of this hierarchy. Gross and Tucker [13] demonstrated that "stratified graphs," which are further up in the hierarchy, are actually a complete invariant of homeomorphism type. Chen and Gross [1]–[5] studied the average genus of a graph for potential applicability to isomorphism testing and showed that a given value of average genus is shared by at most finitely many nonhomeomorphic cutedge-free graphs. Therefore, the average genus of graphs can be a possible criterion, perhaps combined with some other graph invariants, for graph isomorphism candidacy.

The present paper describes an initial effort at combining topological invariants with combinatorial analysis to design efficient graph isomorphism algorithms. We consider the class of graphs whose average genus is bounded by a fixed constant. Interesting graph classes of bounded average genus include trees, cactuses, necklaces, and many others. In particular, there are infinitely many nonplanar graphs that have bounded average genus. In fact, for each fixed integer g, there are infinitely many graphs of average genus less than $2g$ whose minimum genus equals g.

By strengthening the techniques recently developed in topological graph theory,

[*] Received by the editors March 18, 1991; accepted for publication (in revised form) July 16, 1993. This work was partially supported by National Science Foundation grant CCR-9110824. An extended abstract of this work was reported in *The 18th International Workshop on Graph-Theoretic Concepts in Computer Science* (*WG'92*), Lecture Notes in Computer Science, 657 (1993), pp. 103–113.

[†] Department of Computer Science, Texas A&M University, College Station, Texas 77843-3112.

specifically, those in [2]–[5], we are able to sharpen some of the results in those papers (occasional corollaries are repeated when they follow immediately from the sharpened theorems). These sharpened results are then used to prove that all graphs of bounded average genus share only finitely many nonhomeomorphic "frames." Therefore, two graphs of bounded average genus are isomorphic if and only if their frames are isomorphic and the distributions of their remaining edges on the frames are identical. This leads to a linear-time algorithm for isomorphism of graphs of bounded average genus.

We should remark that a polynomial-time algorithm for isomorphism of graphs of bounded average genus is implied by the algorithm of Filotti and Mayer [6]. Indeed, the average genus of a graph is at least as large as its minimum genus. However, Filotti and Mayer's algorithm runs in time $O(n^{c\gamma_{\min}})$, where $c > 300$ is a constant, and the degree rises with the minimum genus $\gamma_{\min}$. On the other hand, in the time complexity of our algorithm, only the multiplicative coefficient increases with the value of the average genus, and the algorithm always runs in linear time.

The paper consists of five sections. Section 2 contains definitions and some basic results in topological graph theory. Section 3 proves several lemmas on the relationship of average genera of a cutedge-free graph and its subgraphs. Section 4 discusses the structure of cutedge-free graphs whose average genus is bounded. A linear-time algorithm is presented in §5 for testing isomorphism of graphs of bounded average genus.

2. Terminology and preliminaries. It is assumed that the reader is somewhat familiar with topological graph theory, and we briefly review the fundamentals. For further description, see Gross and Tucker [12] or White [20].

A *graph* may have multiple adjacencies or self-adjacencies. An edge e of a graph G is a *cutedge* if the removal of e will disconnect the graph. A graph is *cutedge-free* if it contains no cutedges. An *imbedding* of a graph on a surface must have the "cellularity property" that the interior of every region is simply connected. The closed orientable surface of genus n is denoted S_n.

A *rotation* at a vertex v is a cyclic permutation of the edge-ends incident on v. Thus, a d-valent vertex admits $(d-1)!$ rotations. A list of rotations, one for each vertex of the graph, is called a *rotation system.*

An imbedding of a graph G in an orientable surface induces a rotation system, as follows: The rotation at vertex v is the cyclic permutation corresponding to the order in which the edge-ends are traversed in an orientation-preserving tour around v. Conversely, by the Heffter–Edmonds principle, every rotation system induces a unique imbedding of G into an orientable surface. The bijectivity of this correspondence implies that the number of different ways to imbed a graph of valence sequence $d_1, \ldots, d_n$ into a closed, orientable surface is

$$\prod_{i=1}^{n}(d_i - 1)!.$$

Henceforth, we do not distinguish an imbedding of a graph from its corresponding rotation system. Given a rotation system R of a graph, we denote by $\gamma(R)$ the genus of the corresponding imbedding surface and simply call it "the genus of the rotation system R."

For any graph G, if the number of imbeddings in the surface S_k is denoted g_k,

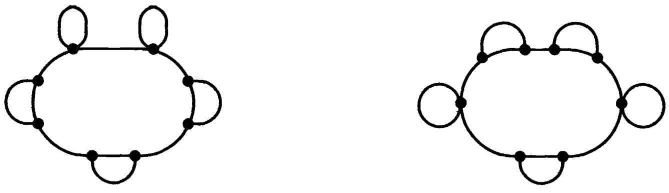


FIG. 1. *Two necklaces of type* (3, 2).

then the sequence

$$g_0, g_1, g_2, \dots$$

is called the *genus distribution* of G. The maximum genus $\gamma_{\max}(G)$ of G is the largest integer r such that $g_r > 0$. The *average genus* of G is defined to be the value

$$\gamma_{\text{avg}}(G) = \frac{\sum_{i=0} i \cdot g_i}{\sum_{i=0} g_i}.$$

It is easy to see that both the minimum genus and the maximum genus of a tree are 0. Therefore, all trees have average genus 0. The following construction, developed by Klein (see Gross, Klein, and Rieper [10]), gives infinitely many cutedge-free graphs whose average genus is less than 1.

A *necklace of type* $(d, 0)$ is a graph $N_{(d,0)} = (V, E)$, such that $V = \{v_1, v_2, \dots, v_{2d}\}$, the vertex v_{2j-1} is connected by a single edge to the vertex v_{2j}, $j = 1, \dots, d$, and the vertex v_{2j-2} is connected by two multiple edges to the vertex v_{2j-1}, where we let v_0 be v_{2d}. A *necklace of type* (d, s) is the necklace $N_{(d,0)}$ of type $(d, 0)$ plus s self-loops added to s distinct interior points of those non-multiple edges of $N_{(d,0)}$ (in an arbitrary way).

The graphs in Fig. 1 are two necklaces of type (3, 2). Although there is a unique necklace of type $(d, 0)$ up to isomorphism, there is more than one nonisomorphic necklace of type (d, s) for $d > 1$ and $s > 1$.

THEOREM 2.1 ([10]). *The average genus of any necklace of type* (r, s) *is* $1 - (\frac{1}{2})^r(\frac{2}{3})^s$.

The technique of *ear adding* is useful in our discussion. We state the result for the technique as follows. The interested reader is referred to Ringeisen [17] for more detailed discussion.

Let G be a graph. A simple path $p = \{u_1, \dots, u_s\}$ is an *ear* to G if its two endpoints u_1 and u_s are vertices of G and none of its interior vertices belong to G. Suppose that we add the ear p to an imbedding $I(G)$ of G to obtain an imbedding $I(G + p)$. There are two different cases.

If the ear p is added to the imbedding $I(G)$ so that its two endpoints u_1 and u_s are inserted between two corners of one face F in $I(G)$, then the ear p splits the face F, and the imbedding $I(G + p)$ has the same genus as that of $I(G)$. Moreover, every vertex of the added ear p has two corners belonging to different faces in the imbedding $I(G + p)$. On the other hand, if the ear p is added to $I(G)$ so that its endpoints u_1 and u_s are inserted between corners of two different faces, then both those faces are merged into one larger face, and the imbedding $I(G + p)$ has genus one larger than that of $I(G)$.

Let G be a subgraph of a graph G'. Denote by $G' - G$ the graph obtained from G' by first deleting all edges in G and then deleting all isolated vertices. A rotation system R of G is an *induced rotation system* of a rotation system R' of G', if R is obtained by deleting all edges in $G' - G$ (and then deleting all isolated vertices) from the rotation system R'. Let Γ'_R be the set of rotation systems of G' whose induced rotation system of G is R, and let Γ and Γ' be the sets of rotation systems of G and G', respectively.

THEOREM 2.2. *The set Γ' is a disjoint union of the sets Γ'_R over all rotation systems R of G. Moreover, $|\Gamma'| = |\Gamma| \cdot |\Gamma'_R|$ for any rotation system R of G.*

Proof. Since each rotation system R' of G' has a unique induced rotation system of G, two sets Γ'_{R_1} and Γ'_{R_2} are disjoint for two different rotation systems R_1 and R_2 of G. Therefore, the set Γ' is a disjoint union of the sets Γ'_R over all rotation systems R of G.

Now let $V_1 = \{u_1, \ldots, u_s\}$ be the set of vertices of G' that are in both G and $G' - G$, and let $V_2 = \{v_1, \ldots, v_t\}$ be the set of vertices of G' that are not in G. Then the number of rotation systems of G' in the set Γ'_R, which is equal to the number of different ways to construct a rotation system of G' from R, depends only on the valences of the vertices of V_1 in the graphs G and $G' - G$ and the valence of the vertices of V_2 in the graph $G' - G$. More precisely, if we denote by $\mathrm{val}_1(v)$ and $\mathrm{val}_2(v)$ the valences of the vertex v in the graphs G and $G' - G$, respectively, then for each $u_k \in V_1$, the number of ways to construct a rotation at u_k for the graph G' based on the rotation at u_k in R is $\prod_{i=0}^{\mathrm{val}_2(u_k)-1}(\mathrm{val}_1(u_k) + i)$, and for each $v_j \in V_2$, the number of ways to construct a rotation at v_j for G' based on R is $(\mathrm{val}_2(v_j) - 1)!$. Consequently, the number of ways to construct a rotation system of G' from R is equal to

$$\left(\prod_{k=1}^{s} \prod_{i=0}^{\mathrm{val}_2(u_k)-1} (\mathrm{val}_1(u_k) + i)\right) \left(\prod_{j=1}^{t} (\mathrm{val}_2(v_j) - 1)!\right),$$

which is a number independent of the rotation system R. Therefore, all sets Γ'_R, for $R \in \Gamma$, contain the same number of rotation systems of G'. Consequently, $|\Gamma'| = |\Gamma| \cdot |\Gamma'_R|$ for every rotation system R of G. □

COROLLARY 2.3. *The average genus of a graph is not less than the average genus of any of its subgraphs.*

Proof. We follow the notations used in Theorem 2.2. It is easy to see that each rotation system of G' in the set Γ'_R has genus at least $\gamma(R)$. By Theorem 2.2,

$$\begin{aligned}\gamma_{\mathrm{avg}}(G') &= \frac{\sum_{R' \in \Gamma'} \gamma(R')}{|\Gamma'|} = \frac{\sum_{R \in \Gamma} \sum_{R' \in \Gamma'_R} \gamma(R')}{|\Gamma'|} \\ &\geq \frac{\sum_{R \in \Gamma} |\Gamma'_R| \gamma(R)}{|\Gamma'|} = \frac{\sum_{R \in \Gamma} \gamma(R)}{|\Gamma|} \\ &= \gamma_{\mathrm{avg}}(G). \quad \square\end{aligned}$$

A graph with two vertices and n edges adjoining them is called a *dipole* and is denoted D_n. A graph with one vertex and n self-loops is called a *bouquet* and is denoted B_n.

COROLLARY 2.4. *A cutedge-free graph that is not a simple cycle has average genus at least $\frac{1}{3}$.*

Proof. Every cutedge-free graph G that is not a simple cycle contains either a subgraph homeomorphic to the bouquet B_2 or a subgraph homeomorphic to the dipole

D_3. It is easy to check that $\gamma_{\text{avg}}(B_2) = \frac{1}{3}$ and $\gamma_{\text{avg}}(D_3) = \frac{1}{2}$. Thus $\gamma_{\text{avg}}(G) \geq \frac{1}{3}$ by Corollary 2.3. □

Given two disjoint graphs G and H, the *bar-amalgamation* of G and H is the result of running a new edge (called the "bar") from a vertex of G to a vertex of H. The following theorem is a direct consequence of Theorem 5 of Gross and Furst [9].

THEOREM 2.5. *The average genus of a bar-amalgamation of two graphs equals the sum of their average genera.*

A maximal cutedge-free subgraph of a connected graph G is called a *cutedge-free component* of the graph G. The following slight generalization of Theorem 2.5 can be easily proved by simple induction.

THEOREM 2.6. *The average genus of a connected graph equals the sum of the average genera of its cutedge-free components.*

3. On the average genera of a graph and its subgraphs. We present a few lemmas concerned with the following problem: Given a graph G and a subgraph H of G with some particular combinatorial property, what can we say about the values of the average genus of them?

The beautiful theory of bridges and attachments, as presented by Tutte [18] in his classic study of graph connectivity, is needed for our discussion in this section. We slightly modify the terminology used in [19].

Let H be a subgraph of a graph G. An *attachment* of H in G is a vertex that lies both in H and in $G - H$.

If some edge of $G - H$ has both its endpoints in the subgraph H, then it is called a *trivial bridge* of $G - H$. Let W be the (possibly not connected) subgraph of G obtained by deleting all vertices (and the edges incident on them) in $V(H)$ from G and let D be a connected component of W. Let B be the subgraph of G obtained from D by adjoining to it each edge of G with one endpoint in D and one in H, together with the end-vertex of that edge in H. Then B is called a *nontrivial bridge* of $G - H$. A subgraph B of G is called a *bridge* of $G - H$ if B is either a trivial bridge or a nontrivial bridge.

Let B be a bridge of $G - H$. Then B is a connected subgraph of G. Moreover, a nontrivial bridge with some of its attachments deleted is also a connected subgraph of G. These follow directly from the definitions. For further discussion of bridges and attachments, see Tutte [19].

A graph is *smooth* if it does not contain 2-valent vertices.

Let H be a subgraph of a graph G. Recall that a simple path P in $G - H$ is an *ear* to H, if the endpoints of P are vertices in H and none of the interior vertices of P are contained in H. An ordered collection $[P_1, P_2, \ldots, P_r]$ of simple paths in $G - H$ is called a *sequence of ears in $G - H$* if, for $1 \leq i \leq r$, the path P_i is an ear to $H + P_1 + \cdots + P_{i-1}$.

LEMMA 3.1. *Let H be a subgraph of a graph G, and let $[P, P']$ be a sequence of ears in $G - H$, such that one endpoint of P' is at an interior vertex of P. Then*

$$\gamma_{\text{avg}}(H + P + P') \geq \gamma_{\text{avg}}(H) + \tfrac{1}{3}.$$

Proof. Let $H' = H + P$, and $H'' = H + P + P'$. Suppose that the two endpoints of P' are v and u, where v is an interior vertex of P.

Given a rotation system R of the graph H, let Γ''_R be the set of rotation systems of the graph H'' whose induced rotation system of H is R. Each rotation system R'' of H'' in the set Γ''_R can be obtained by first attaching the path P to the rotation system R, resulting in a rotation system R' of the graph H', then attaching the path P' to the rotation system R'. There are two different cases.

1. The path P merges two different faces of the rotation system R. Then $\gamma(R') = \gamma(R) + 1$. Hence, no matter how we attach the path P' to the rotation system R', we always have $\gamma(R'') \geq \gamma(R') = \gamma(R) + 1$.

2. The path P splits a face of the rotation system R. Then, attaching the path P to R does not increase the genus. However, the two corners of the vertex v, which is an interior vertex of the path P, belong to different faces of the resulting rotation system R'. Now we attach the path P' to R' by inserting its two endpoints to corners of the vertices v and u in R', respectively. If $v \neq u$, then for each corner c of the vertex u in the rotation system R', at least one of the two corners of the vertex v in R' is contained in a face that does not contain the corner c. Therefore, if the endpoint u of P' is inserted to the corner c, then at least one of the two ways to insert the other endpoint of P' to a corner of v makes P' merge two different faces in R', thus increasing the genus by 1. Consequently, in this case, at least half of the ways to attach the path P' to the rotation system R' increase the genus by 1. In the other case, if $v = u$, then in the six ways to attach the simple cycle P' to the vertex v in R', exactly two of them merge two different faces and thereby increase the genus by 1.

Therefore, for each rotation system R' of H' obtained by attaching the path P to the rotation system R, at least one third of the rotation systems of H'' that are obtained by attaching the path P' to R' have genus at least $\gamma(R) + 1$. Consequently, at least one third of the rotation systems in the set Γ''_R have genus greater than or equal to $\gamma(R) + 1$. Note that no rotation system in the set Γ''_R has genus less than $\gamma(R)$.

Let Γ and Γ'' be the sets of rotation systems of the graphs H and H'', respectively. Then

$$\begin{aligned}
\gamma_{\text{avg}}(H + P + P') = \gamma_{\text{avg}}(H'') &= \frac{\sum_{R''\in\Gamma''} \gamma(R'')}{|\Gamma''|} \\
&= \frac{\sum_{R\in\Gamma} \sum_{R''\in\Gamma''_R} \gamma(R'')}{|\Gamma''|} \\
&\geq \frac{\sum_{R\in\Gamma} (|\Gamma''_R|\gamma(R) + |\Gamma''_R|/3)}{|\Gamma''|} \\
&= \frac{\sum_{R\in\Gamma} (\gamma(R) + 1/3)}{|\Gamma|} \\
&= \gamma_{\text{avg}}(H) + \tfrac{1}{3},
\end{aligned}$$

where we have used the facts that Γ'' is the disjoint union of the sets Γ''_R over all $R \in \Gamma$, and that $|\Gamma''| = |\Gamma''_R| \cdot |\Gamma|$ for every rotation system R of the subgraph H. □

LEMMA 3.2. *Let H be a subgraph of a cutedge-free smooth graph G such that $G - H$ contains a nontrivial bridge. Then there is a sequence of two ears $[P, P']$ in $G - H$ such that*

$$\gamma_{\text{avg}}(H + P + P') \geq \gamma_{\text{avg}}(H) + \tfrac{1}{3}.$$

Proof. Suppose that B is a nontrivial bridge of $G - H$. Since G is cutedge-free, the bridge B contains a simple path P that is an ear to H. The path P cannot be a single edge, since P is putatively contained in the nontrivial bridge B. Moreover, the graph G contains no 2-valent vertices. Therefore, we must be able to find another simple path P' in B from an interior vertex of the path P to another vertex, which is either an interior vertex of P or an attachment of the subgraph H, such that no interior vertex of P' lies on $H + P$. Thus the collection $[P, P']$ is a sequence of ears in $G - H$ and the ear P' has an endpoint at an interior vertex of P.

By Lemma 3.1, we have

$$\gamma_{\text{avg}}(H + P + P') \geq \gamma_{\text{avg}}(H) + \tfrac{1}{3}. \qquad \Box$$

A *suspended chain C* (or simply *a chain*) of a graph G is a simple path in G such that all interior vertices of C have valence 2. A suspended chain C of G is *maximal* if neither of its end-vertices has valence 2.

Let G be a cutedge-free smooth graph and let H be a subgraph of G. Two trivial bridges in $G-H$ *overlap on a chain C of H* if either they share a common attachment that is an interior vertex of C, or one of the bridges has two distinct endpoints p_1 and p_2 on the chain C, and the other bridge has an endpoint q between p_1 and p_2 on the chain C such that $q \neq p_1, p_2$.

LEMMA 3.3. *Let H be a subgraph of a cutedge-free smooth graph G such that $G - H$ contains two trivial bridges B and B' overlapping on a chain C of H. Then*

$$\gamma_{\text{avg}}(H + B + B') \geq \gamma_{\text{avg}}(H) + \tfrac{1}{3}.$$

Proof. We first suppose that the bridge B has two distinct endpoints p_1 and p_2 on the chain C and that the bridge B' has an endpoint q between p_1 and p_2 on the chain C such that $q \neq p_1, p_2$. Consider the graph $H' = H - C' + B$, which is obtained by deleting the subchain C' adjoining p_1 and p_2 on the chain C, then adding the edge B to the resulting graph. It is obvious that the graph H' is homeomorphic to the graph H and $H + B + B' = H' + C' + B'$. Moreover, $[C', B']$ is a sequence of ears in $G - H'$ such that one endpoint q of B' is at an interior vertex of the path C'. By Lemma 3.1, we have

$$\begin{aligned}\gamma_{\text{avg}}(H + B + B') &= \gamma_{\text{avg}}(H' + C' + B')\\ &\geq \gamma_{\text{avg}}(H') + \tfrac{1}{3} = \gamma_{\text{avg}}(H) + \tfrac{1}{3}.\end{aligned}$$

For the other case, suppose that the two trivial bridges B and B' share a common attachment a that is 2-valent in H. There are three possible subcases:

1. Neither of the bridges B and B' is a self-loop;
2. Exactly one of the bridges B and B' is a self-loop;
3. Both the bridges B and B' are self-loops.

To complete the proof, it suffices to show that in each of these three subcases, adding the bridges B and B' to the subgraph H raises its average genus by at least $\frac{1}{3}$. The considerations are similar for all three subcases, and we presently provide details for subcase 3.

Subcase 3. We consider all the ways to extend an arbitrary imbedding of H by adding first B and then B', and we see that at least one third of them increase the genus. If the vertex a lies on two faces of the imbedding of H, then two of the six ways to attach the self-loop B already increase the genus, thereby assuring that the genus increases for at least one third of the ways to add both B and B'. Therefore, let us suppose that the vertex a lies on only one face of the imbedding of H, so that none of the six ways of adding self-loop B increases the genus. In four of these six

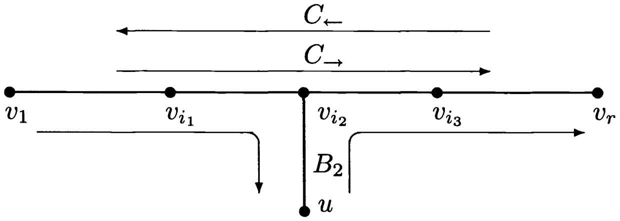


FIG. 2. *The edge B_2 splits a face in R_2.*

imbeddings of $H + B$, the self-loop B bounds a monogon, and six of the 20 ways to add self-loop B' increase the genus. In the other two imbeddings of $H + B$, eight of the 20 ways to add self-loop B' increase the genus. Thus, of the 120 ways to add both B and B', 40 (i.e., $4 \cdot 6 + 2 \cdot 8$) increase the genus. □

A trivial bridge B in $G - H$ is a *T-hanging ear* on a maximal chain C of the subgraph H if one endpoint of B is at an interior vertex of C, while the other endpoint of B is at a vertex of H that is not an interior vertex of C.

LEMMA 3.4. *Let H be a subgraph of a graph G. Suppose that B_1, B_2, and B_3 are three T-hanging ears on a maximal chain C of H such that no two of them overlap on C. Then*

$$\gamma_{\text{avg}}(H + B_1 + B_2 + B_3) \geq \gamma_{\text{avg}}(H) + \tfrac{1}{2}.$$

Proof. Suppose that $C = v_1 v_2 \ldots v_r$. Since the bridges B_1, B_2, and B_3 do not overlap, we can suppose, without loss of generality, that v_{i_1}, v_{i_2}, and v_{i_3} are the three endpoints of B_1, B_2, and B_3 on the chain C, respectively, with $1 < i_1 < i_2 < i_3 < r$. Let $H_2 = H{+}B_2$, $H_{12} = H{+}B_1{+}B_2$, $H_{23} = H{+}B_2{+}B_3$, and $H_{123} = H{+}B_1{+}B_2{+}B_3$.

Let Γ_R denote the set of rotation systems of H_{123} whose induced rotation system of H is R. Every rotation system in the set Γ_R can be obtained by attaching the three edges B_1, B_2, and B_3 to the rotation system R.

For any rotation system R of H, since every interior vertex of C is 2-valent in H, the chain C gives rise to two directed sides

$$C_{\rightarrow} = (v_1 \ldots v_{i_1} \ldots v_{i_2} \ldots v_{i_3} \ldots v_r),$$
$$C_{\leftarrow} = (v_r \ldots v_{i_3} \ldots v_{i_2} \ldots v_{i_1} \ldots v_1),$$

such that each of the directed sides is a subwalk of the boundary of a face in the rotation system R. Now attaching the edge B_2 to R results in a rotation system R_2 of H_2. If B_2 merges two different faces in R, then $\gamma(R_2) = \gamma(R) + 1$. Therefore, an arbitrary way of attaching the edges B_1 and B_3 to R_2 results in a rotation system R_{123} of H_{123}, such that $\gamma(R_{123}) \geq \gamma(R_2) = \gamma(R) + 1$.

For the other case, suppose that B_2 splits a face in R. Without loss of generality, we suppose that one endpoint of B_2 is inserted to v_{i_2} from the directed side $C_{\rightarrow}$ and the other endpoint of B_2 is u. See Fig. 2. Since the two sides of B_2 appear in different faces of the resulting rotation system R_2 of H_2, the two walks $v_1 \ldots v_{i_1} \ldots v_{i_2} u$ and $u v_{i_2} \ldots v_{i_3} \ldots v_r$ are subwalks of boundaries of different faces of R_2. Consequently, in the rotation system R_2, there are a corner c_1 of the vertex v_{i_1} and a corner c_3 of the vertex v_{i_3} such that c_1 and c_3 belong to different faces. On the other hand, since the bridge B_2 is a T-hanging ear on the chain C, the vertex u is not an interior vertex of the chain C. Therefore, the other directed side $C_{\leftarrow} = (v_r \ldots v_{i_3} \ldots v_{i_2} \ldots v_{i_1} \ldots v_1)$ of the chain C is still a subwalk of a face boundary in the rotation system R_2. Thus the other corner c_1' of the vertex v_{i_1} and the other corner c_3' of the vertex v_{i_3} belong to the same face. Therefore, at least one of the vertices v_{i_1} and v_{i_3} has its two corners belonging to different faces in the rotation system R_2. Without loss of generality, suppose that the two corners c_1 and c_1' of v_{i_1} belong to different faces in R_2 (the case that the two corners of v_{i_3} belong to different faces can be proved similarly). Then, as we analyzed in the proof of Lemma 3.1, at least half of the ways to attach B_1 to R_2 increase the genus by 1. Because arbitrarily attaching the edge B_3 then does not decrease the genus, we conclude that at least half of the ways to attach the edges B_1 and B_3 to the rotation system R_2 result in a rotation system R_{123} of H_{123} satisfying

$$\gamma(R_{123}) \geq \gamma(R_2) + 1 = \gamma(R) + 1.$$

Therefore, at least half of the rotation systems in the set Γ_R have genus at least $\gamma(R) + 1$ (and no rotation systems in Γ_R have genus less than $\gamma(R)$).

Now, by a calculation completely similar to that in the proof of Lemma 3.1, we are able to conclude that

$$\gamma_{\text{avg}}(H_{123}) \geq \gamma_{\text{avg}}(H) + \tfrac{1}{2}. \qquad \square$$

4. Cutedge-free graphs of bounded average genus. We recall that a dipole D_n is a graph with two vertices and n edges adjoining them, and a bouquet B_n is a graph with one vertex and n self-loops.

LEMMA 4.1. *For every integer $n \geq 3$, $\gamma_{\text{avg}}(D_n) \geq \frac{n}{6}$.*

Proof. In Chen and Gross [2], it was proved that

$$\gamma_{\text{avg}}(D_n) = \frac{n}{2} - \left(\frac{H_{n+1}}{2} + (-1)^n \frac{1}{2n(n+1)} \right)$$

for every integer $n \geq 1$, where H_n is the *harmonic number* defined by $H_n = \sum_{k=1}^n \frac{1}{k}$. For the values $3 \leq n \leq 6$, it can be easily checked that $\gamma_{\text{avg}}(D_n) \geq \frac{n}{6}$. Moreover, since, for $n \geq 7$, we have $\frac{n}{4} \geq H_{n+1}/2 + (-1)^n/(2n(n+1))$, thus $\gamma_{\text{avg}}(D_n) \geq \frac{n}{4}$ for $n \geq 7$. $\square$

LEMMA 4.2. *For every integer $n \geq 2$, $\gamma_{\text{avg}}(B_n) \geq n/6$.*

Proof. Let $g_m(n)$ be the number of imbeddings of the bouquet B_n in the surface S_m. Gross, Robbins, and Tucker [11] have shown that

$$(1) \qquad \begin{aligned} (n+1)g_m(n) = 4(2n-1)(2n-3)(n-1)^2(n-2)g_{m-1}(n-2) \\ +4(2n-1)(n-1)g_m(n-1) \end{aligned}$$

with the boundary conditions

$$\begin{gathered} g_m(n) = 0 \quad \text{if } m < 0 \text{ or } n < 0, \\ g_0(0) = g_0(1) = 1 \quad \text{and} \quad g_m(0) = g_m(1) = 0 \quad \text{for } m > 0, \\ g_0(2) = 4, \quad g_1(2) = 2, \quad g_m(2) = 0 \quad \text{for } m > 1. \end{gathered}$$

Moreover, by Xuong's theorem on the maximum genus of a graph [21], it is easy to see that $\gamma_{\max}(B_n) = \gamma_{\max}(B_{n-2}) + 1$ for all $n \geq 3$.

From (1), we have

$$(n+1)g_0(n) = 4(2n-1)(n-1)g_0(n-1).$$

Thus it is easy to check, by induction, that

$$(2) \qquad g_0(n) = \frac{2^n(2n-1)!}{(n+1)!} \leq \frac{(2n-1)!}{3} \quad \text{for } n \geq 3.$$

Now we prove $\gamma_{\text{avg}}(B_n) \geq \frac{n}{6}$ for $n \geq 2$. The inequality can be easily checked for the case where $n = 2, 3$, so we suppose that $n \geq 4$. We have

$$\gamma_{\text{avg}}(B_n) = \frac{\sum_{i=1}^{M(n)} i g_i(n)}{(2n-1)!} = \frac{\sum_{i=1}^{M(n)} (i-1) g_i(n) + \sum_{i=1}^{M(n)} g_i(n)}{(2n-1)!},$$

where we have abbreviated $\gamma_{\max}(B_n)$ by $M(n)$. By (2), we have

$$\frac{\sum_{i=1}^{M(n)} g_i(n)}{(2n-1)!} = 1 - \frac{g_0(n)}{(2n-1)!} \geq \frac{2}{3}.$$

Moreover, by (1), we have

$$(n+1) g_m(n) \geq 4(2n-1)(2n-3)(n-1)^2(n-2) g_{m-1}(n-2).$$

Therefore,

$$\begin{aligned}
\gamma_{\text{avg}}(B_n) &\geq \frac{4(2n-1)(2n-3)(n-1)^2(n-2)\sum_{i=1}^{M(n)}(i-1)g_{i-1}(n-2)}{(n+1)(2n-1)!} + \frac{2}{3} \\
&= \frac{4(2n-1)(2n-3)(n-1)^2(n-2)\sum_{i=0}^{M(n-2)} i g_i(n-2)}{(n+1)(2n-1)!} + \frac{2}{3} \\
&= \frac{n-1}{n+1}\gamma_{\text{avg}}(B_{n-2}) + \frac{2}{3}.
\end{aligned}$$

Here we used the facts that $M(n) = M(n-2) + 1$ and that the number of rotation systems of B_{n-2} is $(2n-5)!$. By the inductive hypothesis, $\gamma_{\text{avg}}(B_{n-2}) \geq (n-2)/6$. Therefore,

$$\gamma_{\text{avg}}(B_n) \geq \frac{(n-1)(n-2)}{6(n+1)} + \frac{2}{3} > \frac{n}{6}. \qquad \square$$

We remark that Chen and Gross [2] showed that for every sufficiently large number n, $\gamma_{\text{avg}}(D_n) \geq \frac{n}{4}$ and $\gamma_{\text{avg}}(B_n) \geq \frac{n}{4}$.

In the rest of this paper, we let $\Delta > 0$ be a fixed real number. Denote by $\lceil \Delta \rceil$ the smallest integer that is not less than Δ.

LEMMA 4.3. *Let G be a cutedge-free smooth graph of average genus less than Δ. Then G has a cutedge-free subgraph H with the following properties*: (1) *H has at most* $12\lceil \Delta \rceil$ *vertices of valence greater than* 2, *and at most* $18\lceil \Delta \rceil$ *different maximal chains*; (2) *all bridges of $G-H$ are trivial bridges*; (3) *no two trivial bridges of $G-H$ overlap on a chain of H*; *and* (4) *no three T-hanging ears are on a single maximal chain of H.*

Proof. We begin by selecting a subgraph H_0 of G that is a simple cycle. Then we construct a sequence of cutedge-free subgraphs H_i, $i = 0, 1, 2, \ldots$, of G inductively. Suppose that we have obtained a subgraph H_i of G. We augment H_i in any one of the following three ways, if any of them is applicable, to produce the subgraph H_{i+1}.

1. If there is a nontrivial bridge in $G - H_i$, then by Lemma 3.2, we can find a sequence $[P_i, P_i']$ of two ears in $G - H_i$ such that P_i' has an endpoint on an interior vertex of P_i. We let $H_{i+1} = H_i + P_i + P_i'$.

2. If there are two trivial bridges B_i and B_i' in $G - H_i$ that overlap on a chain of H_i, then we let $H_{i+1} = H_i + B_i + B_i'$.

3. If there are three T-hanging ears B_i, B_i', and B_i'' in $G - H_i$ on a maximal chain C of H_i such that no two of them overlap on C, then we let $H_{i+1} = H_i + B_i + B_i' + B_i''$.

Since each H_i constructed above is a subgraph of the graph G, it follows from Corollary 2.3 that the average genus of H_i is less than Δ. Moreover, by Lemma 3.1, Lemma 3.3, and Lemma 3.4, the increment in average genus from H_i to H_{i+1} is at least $\frac{1}{3}$. Therefore, there is an index I, $I < 3\lceil\Delta\rceil$, such that none of the above conditions apply to the graph H_I.

Now we count the number of vertices of valence greater than 2 and the number of different maximal chains in H_I. The graph H_I is obtained by adding a sequence S of ears in $G - H_0$ to the simple cycle H_0. In case 1 and case 2, adding two ears increases the average genus by at least $\frac{1}{3}$ (Lemma 3.1 and Lemma 3.3), and in case 3, adding three ears increases the average genus by at least $\frac{1}{2}$ (Lemma 3.4). Therefore, on average, adding an ear in the sequence S increases the average genus by at least $\frac{1}{6}$. Consequently, the sequence S consists of at most $6\lceil\Delta\rceil$ ears, since the average genus of the graph H_I is less than Δ. Because adding an ear to a graph increases the number of vertices of valence greater than 2 by at most 2 and increases the number of maximal chains by at most 3, and the graph H_I is obtained by adding at most $6\lceil\Delta\rceil$ ears to the simple cycle H_0, we conclude that the graph H_I contains at most $2 \cdot 6\lceil\Delta\rceil = 12\lceil\Delta\rceil$ vertices of valence greater than 2 and at most $3 \cdot 6\lceil\Delta\rceil = 18\lceil\Delta\rceil$ different maximal chains. □

Two multiple edges e_1 and e_2 in a graph G are called a *pair of twin multiple edges*, if e_1 and e_2 are the only edges adjoining two distinct 3-valent vertices in G.

Let G be a graph and let H be a cutedge-free subgraph of G such that all bridges in $G - H$ are trivial bridges. A trivial bridge e in $G - H$ is a *U-hanging ear* on a maximal chain C in H, if e does not overlap any other edges in $G - H$ on C, and either the two endpoints of e are identical and located at an interior vertex of C, or the two endpoints of e are located at two adjacent interior vertices of C. We observe that a U-hanging ear in $G - H$ is either a self-loop located at a vertex that has valence 4 in the graph G, or an edge in a pair of twin multiple edges in the graph G.

DEFINITION 4.4. *Let G be a cutedge-free graph. A* frame $F(G)$ *of G is a graph obtained in the following way:*

1. *If e is a self-loop in G whose endpoint is of valence 4, then delete e. Reiterate until no such self-loops remain;*

2. *For each pair of twin multiple edges in G, delete either one of the twin edges. Reiterate until no twin multiple edges remain.*

In other words, a frame $F(G)$ of G is obtained from G by deleting a maximum set of U-hanging ears. We observe that a frame of G is not uniquely defined since, for a pair of twin multiple edges, either of them can be contained in a frame of G. However, all frames of G are homeomorphic and have the same number of different maximal chains.

THEOREM 4.5. *Let G be a cutedge-free smooth graph of average genus less than Δ. Then any frame $F(G)$ of G contains at most $(8\lceil\Delta\rceil)^3$ different maximal chains.*

Proof. By Lemma 4.3, we can find a cutedge-free subgraph H of G, which has $\alpha \leq 12\lceil\Delta\rceil$ vertices of valence greater than 2 and $\beta \leq 18\lceil\Delta\rceil$ different maximal chains, such that all bridges in $G - H$ are trivial bridges, no two trivial bridges in $G - H$ overlap on a chain of H, and no three T-hanging ears are on a single maximal chain of H. Therefore, every edge e in $G - H$ must lie in one of the following four possible classes:

1. e is a T-hanging ear on a maximal chain in H;
2. e is a self-loop at a vertex of valence greater than 2 in H;
3. e has its two endpoints at two different vertices of valence greater than 2 in H;
4. e is a U-hanging ear in $G - H$.

We now consider how many edges in $G - H$ can belong to classes 1–3.

By the construction of the graph H, there are at most two T-hanging ears on each maximal chain in H. Therefore, there are at most $2 \cdot \beta \leq 36\lceil\Delta\rceil$ edges in $G - H$ that belong to class 1.

Let the vertices of valence greater than 2 in H be u_1, u_2, …, u_α. Suppose that there are $h(u_i)$ self-loops in $G-H$ located at the vertex u_i. Without loss of generality, suppose that $h(u_i) \geq 2$ for $i = 1, 2, \ldots, \alpha'$, and $h(u_j) \leq 1$ for $j = \alpha' + 1, \ldots, \alpha$. Consider the subgraph G' that consists of a spanning tree of G and the self-loops in $G - H$ located at vertices of valence greater than 2 in H. By Theorem 2.6, the average genus of the graph G' is the sum of the average genera of the α bouquets $B_{h(u_1)}$, $B_{h(u_2)}$, …, $B_{h(u_\alpha)}$, where the bouquet $B_{h(u_i)}$ consists of the vertex u_i and the $h(u_i)$ self-loops in $G - H$ that are located at u_i. Therefore, we have

$$\begin{aligned}\gamma_{\mathrm{avg}}(G') &= \gamma_{\mathrm{avg}}(B_{h(u_1)}) + \gamma_{\mathrm{avg}}(B_{h(u_2)}) + \cdots + \gamma_{\mathrm{avg}}(B_{h(u_\alpha)}) \\ &\geq \gamma_{\mathrm{avg}}(B_{h(u_1)}) + \gamma_{\mathrm{avg}}(B_{h(u_2)}) + \cdots + \gamma_{\mathrm{avg}}(B_{h(u_{\alpha'})}) \\ &\geq \frac{h(u_1) + h(u_2) + \cdots + h(u_{\alpha'})}{6},\end{aligned}$$

where the result of Lemma 4.2 that $\gamma_{\mathrm{avg}}(B_n) \geq \frac{n}{6}$ for $n \geq 2$ has been used. Now since G' is a subgraph of G, by Corollary 2.3, $\gamma_{\mathrm{avg}}(G') < \Delta$. Therefore, the total number of edges in class 2 is (note $h(u_j) \leq 1$ for $j = \alpha' + 1, \ldots, \alpha$)

$$\begin{aligned}&h(u_1) + \cdots + h(u_{\alpha'}) + h(u_{\alpha'+1}) + \cdots + h(u_\alpha) \\ &\leq 6\gamma_{\mathrm{avg}}(G') + h(u_{\alpha'+1}) + \cdots + h(u_\alpha) \\ &\leq 6\Delta + (\alpha - \alpha') \\ &\leq 6\Delta + \alpha \\ &\leq 18\lceil\Delta\rceil.\end{aligned}$$

Finally, for each pair of vertices of valence greater than 2 in the subgraph H, the number h of multiple edges that can be attached is at most $6\lceil\Delta\rceil$. Otherwise, what would result is a supergraph G'' of the dipole D_h, whose average genus is, by Lemma 4.1, at least $\frac{h}{6}$, which is greater then $\lceil\Delta\rceil$. This would contradict Corollary 2.3, since the graph G'' is a subgraph of the graph G. Therefore, there are at most $6\lceil\Delta\rceil\binom{\alpha}{2} \leq 432\lceil\Delta\rceil^3 - 36\lceil\Delta\rceil^2$ edges in class 3.

Let H' be the subgraph of G obtained by adding to the graph H all edges in $G-H$ that are in classes 1–3. We calculate the number of maximal chains in H'. The graph H' is obtained by adding at most $36\lceil\Delta\rceil$ edges in class 1, at most $18\lceil\Delta\rceil$ edges in class 2, and at most $432\lceil\Delta\rceil^3 - 36\lceil\Delta\rceil^2$ edges in class 3. Since there are at most two T-hanging ears on each maximal chain of H, at most two new vertices of valence greater than 2 are created on each maximal chain of H, thus each maximal chain of H is subdivided into at most three maximal chains in the graph H'. Moreover, each edge in classes 1–3 becomes a maximal chain in H'. Therefore, the number of different maximal chains in the graph H' is less than or equal to

$$3\beta + 36\lceil\Delta\rceil + 18\lceil\Delta\rceil + (432\lceil\Delta\rceil^3 - 36\lceil\Delta\rceil^2) \leq 504\lceil\Delta\rceil^3 < (8\lceil\Delta\rceil)^3.$$

$\Delta > 0$, thus $\lceil\Delta\rceil \geq 1$ is used here.

By the construction of the graph H', all edges in $G - H'$ are U-hanging ears in $G - H$. Since a U-hanging ear e in $G - H$ does not overlap any other edge in $G - H$, it follows that the edge e is also a U-hanging ear in $G - H'$. Therefore, all edges in $G - H'$ are U-hanging ears in $G - H'$. Since a frame of G is a subgraph of G obtained by deleting a maximum set of U-hanging ears, we conclude that the graph H' contains a subgraph F that is a frame of G. Thus the number of different maximal chains of F is also less than $(8\lceil\Delta\rceil)^3$. The theorem is proved, because all frames of G have the same number of different maximal chains. □

5. The algorithm. We present in this section a linear-time algorithm for isomorphism of graphs of bounded average genus.

Given a cutedge-free smooth graph G, let $F(G)$ be a frame of G. All edges in $G - F(G)$ are U-hanging ears. There are two kinds of U-hanging ears: self-loops and edges with two distinct endpoints (they are called *closed ears* and *open ears*, respectively). Since no two U-hanging ears in $G - F(G)$ overlap on a maximal chain of $F(G)$, the U-hanging ears on a maximal chain C of $F(G)$ can be ordered $(e_1, e_2, \ldots, e_r)$ in such a way that if we travel the maximal chain C from one end-vertex to the other end-vertex, we encounter the endpoints of e_1 first, then the endpoints of $e_2, \ldots,$ and finally the endpoints of e_r. Call the ordered sequence $(x(e_1), x(e_2), \ldots, x(e_r))$ the *distribution of U-hanging ears* on the maximal chain C, where $x(e_i) =$ "open," if e_i is an open ear, and $x(e_i) =$ "closed," if e_i is a closed ear.

Let G_1 and G_2 be two graphs, and let ϕ_F be an isomorphism from a frame $F(G_1)$ of G_1 to a frame $F(G_2)$ of G_2, such that for every maximal chain C of $F(G_1)$, the distribution of U-hanging ears on C is identical to the distribution of U-hanging ears on $\phi_F(C)$. Then, obviously the isomorphism ϕ_F can be extended to an isomorphism ϕ from G_1 to G_2 in the following way: $\phi(e) = \phi_F(e)$ for every edge e in $F(G_1)$. For each maximal chain C of $F(G_1)$, if the distributions of U-hanging ears on C and $\phi_F(C)$ are

$$(x(e_1), x(e_2), \ldots, x(e_r)) \quad \text{and} \quad (x(e_1'), x(e_2'), \ldots, x(e_r')),$$

respectively, where $x(e_i) = x(e_i')$ for $i = 1, \ldots, r$, then let $\phi(e_i) = e_i'$, $i = 1, \ldots, r$. The isomorphism ϕ from G_1 to G_2 is said to be *induced* by the isomorphism ϕ_F from $F(G_1)$ to $F(G_2)$. It is easy to see that an isomorphism ϕ from G_1 to G_2 is induced by an isomorphism ϕ_F from $F(G_1)$ to $F(G_2)$ if and only if ϕ_F is the restriction of ϕ on $F(G_1)$.

THEOREM 5.1. *Let $F(G_1)$ and $F(G_2)$ be two arbitrary frames of two graphs G_1 and G_2, respectively. If the graphs G_1 and G_2 are isomorphic, then there is an isomorphism from G_1 to G_2 that is induced by an isomorphism from $F(G_1)$ to $F(G_2)$.*

Proof. Let ϕ be an isomorphism from G_1 to G_2. We construct an isomorphism ϕ' from G_1 and G_2, such that ϕ' is induced by an isomorphism from $F(G_1)$ to $F(G_2)$.

For any edge e of G_1 such that either $e \in F(G_1)$ and $\phi(e) \in F(G_2)$, or $e \notin F(G_1)$ and $\phi(e) \notin F(G_2)$, we let $\phi'(e) = \phi(e)$. If, for an edge e of G_1, either $e \in F(G_1)$ and $\phi(e) \notin F(G_2)$, or $e \notin F(G_1)$ and $\phi(e) \in F(G_2)$, then e must be an edge in a pair of twin multiple edges of G_1. Let e' be the partner of e in G_1. Then we also have either $e' \notin F(G_1)$ and $\phi(e') \in F(G_2)$, or $e' \in F(G_1)$ and $\phi(e') \notin F(G_2)$. So we let $\phi'(e) = \phi(e')$ and $\phi'(e') = \phi(e)$.

It is easy to see that ϕ' is also an isomorphism from G_1 and G_2. Moreover, by the construction of ϕ', an edge e in G_1 is in $F(G_1)$ if and only if $\phi'(e)$ is in $F(G_2)$. Therefore, the restriction ϕ'_F of ϕ' on the frame $F(G_1)$ is an isomorphism from $F(G_1)$ to $F(G_2)$. Consequently, ϕ' is an isomorphism from G_1 to G_2 that is induced by the isomorphism ϕ'_F from $F(G_1)$ to $F(G_2)$. □

Let G_1 and G_2 be two graphs of average genus less than Δ. We first suppose that both graphs G_1 and G_2 are cutedge free and smooth. The following algorithm is used to test the isomorphism of G_1 and G_2.

ALGORITHM 1.

```
1. Construct two arbitrary frames $F(G_1)$ and $F(G_2)$ of $G_1$ and $G_2$, respectively.
2. For each isomorphism $\phi_F$ from $F(G_1)$ to $F(G_2)$, do step 3.
3. If, for every maximal chain $C$ of $F(G_1)$, the distribution of $U$-hanging ears
   on $C$ equals the distribution of $U$-hanging ears on $\phi_F(C)$, then the graphs $G_1$
   and $G_2$ are isomorphic. Construct the isomorphism from $G_1$ to $G_2$ induced
   by the isomorphism $\phi_F$.
```

We remark that Algorithm 1 not only tests the isomorphism of two given graphs, but also lists all isomorphisms of the two graphs that are induced by isomorphisms of the frames $F(G_1)$ and $F(G_2)$.

By Theorem 5.1, if G_1 and G_2 are isomorphic, then there is an isomorphism from G_1 to G_2 that is induced by an isomorphism from $F(G_1)$ to $F(G_2)$. Conversely, if there is an isomorphism of G_1 and G_2 that is induced by an isomorphism of the frames $F(G_1)$ and $F(G_2)$, then of course G_1 and G_2 are isomorphic. Therefore, Algorithm 1 correctly tests the isomorphism of two cutedge-free smooth graphs G_1 and G_2.

Now we discuss the time complexity of the algorithm. The frames $F(G_1)$ and $F(G_2)$ can easily be constructed in linear time, assuming a reasonable encoding of the graphs. Moreover, given an isomorphism ϕ_F from $F(G_1)$ to $F(G_2)$, it is possible in linear time to check for every maximal chain C of $F(G_1)$ that the distribution of U-hanging ears on C is identical to the distribution of U-hanging ears on $\phi_F(C)$, and then to construct the induced isomorphism from G_1 to G_2. Finally, since the graphs G_1 and G_2 have average genus less than Δ, it follows from Theorem 4.5 that each of the frames $F(G_1)$ and $F(G_2)$ has at most $(8\lceil\Delta\rceil)^3$ different maximal chains. Since an isomorphism from $F(G_1)$ to $F(G_2)$ must map a maximal chain of $F(G_1)$ to a maximal chain of $F(G_2)$, we can list all isomorphisms from $F(G_1)$ to $F(G_2)$ by systematically enumerating all possible matchings from the maximal chains of $F(G_1)$ to the maximal chains of $F(G_2)$. Since there are at most $((8\lceil\Delta\rceil)^3)!$ different such matchings, step 3 is executed at most $(8\lceil\Delta\rceil)^3)!$ times. In conclusion, Algorithm 1 runs in linear time.

Simple modifications on Algorithm 1 result in a linear-time algorithm for isomorphism of cutedge-free graphs, which are not necessarily smooth. We sketch the modifications here and leave the details to the interested reader.

Given two cutedge-free graphs G_1 and G_2, we first construct two smooth graphs W_1 and W_2 that are obtained from G_1 and G_2, respectively, by smoothing all 2-valent vertices. Moreover, we also assign a "weight" to each edge in W_1 and W_2, which equals the number of interior vertices of the corresponding maximal chain in G_1 and G_2. It is easy to see that the graphs G_1 and G_2 are isomorphic if and only if there is an isomorphism from W_1 to W_2, regarded as unweighted graphs, such that every edge in

W_1 is mapped to an edge in W_2 with the same weight.

To construct a frame $F(W)$ of a weighted graph W, we delete all self-loops in W that are located at a 4-valent vertex of W, and for each pair of twin multiple edges in W, we delete the one with the *smaller* weight (if the two twin edges have the same weight, delete either one).

The distribution of U-hanging ears in $W - F(W)$ on a maximal chain C of $F(W)$ is defined to be an ordered list

$$\{x(e_1), w(e_1), x(e_2), w(e_2), \ldots, x(e_r), w(e_r)\},$$

where $x(e_i) =$ "open" or "closed" depending on whether e_i is an open or a closed ear, and $w(e_i)$ is the weight of the edge e_i such that, if we travel the maximal chain C from one end-vertex to the other end-vertex, we encounter the endpoints of the edge e_1 first, then the endpoints of the edge e_2, and so on.

With these modifications, it is easy to see that Algorithm 1 still runs in linear time, and tests the isomorphism of cutedge-free, smooth, and weighted graphs of average genus less than Δ. It thereby tests the isomorphism of cutedge-free graphs of average genus less than Δ.

Finally, we discuss the case in which the graphs G_1 and G_2 are not necessarily cutedge free.

By Theorem 2.6, the average genus of a graph G equals the sum of the average genera of its cutedge-free components. Therefore, if the average genus of G is less than Δ, then the average genus of every cutedge-free component of G is also less than Δ. Moreover, if a cutedge-free component D of G is not a simple cycle (call D a *noncycle cutedge-free component* of G), then, by Corollary 2.4, the average genus of D is at least $\frac{1}{3}$. Consequently, the graph G contains at most $3\lceil\Delta\rceil$ noncycle cutedge-free components.

The idea is to replace each noncycle cutedge-free component of G_1 and G_2 by a planar graph, so that G_1 and G_2 are isomorphic if and only if the two resulting planar graphs are isomorphic.

A *wheel* $W(c; v_1, \ldots, v_s)$ consists of an s-cycle $C_s = (v_1, \ldots, v_s)$, together with an extra vertex c joined to the s vertices of C_s by s edges. The vertex c is called the *center* of the wheel. A vertex v of a cutedge-free component D of a graph G is called an *outer vertex* of D, if v is an endpoint of some cutedge of G.

Let G_1 and G_2 be two graphs of average genus less than Δ.

ALGORITHM 2.

1. Find all noncycle cutedge-free components of G_1 and G_2. If the number of noncycle cutedge-free components of G_1 is not equal to that of G_2, G_1 and G_2 are not isomorphic. Otherwise, let

$$D_1, D_2, \ldots, D_k \quad \text{and} \quad D'_1, D'_2, \ldots, D'_k$$

be the noncycle cutedge-free components of G_1 and G_2, respectively. Construct a frame $F(D_i)$ for D_i and a frame $F(D'_i)$ for D'_i, $i = 1, \ldots, k$.

2. For each permutation π of $(1, \ldots, k)$ and each list $(\phi_1, \phi_2, \ldots, \phi_k)$, where ϕ_i is an isomorphism from $D_{\pi(i)}$ to D'_i induced by an isomorphism from $F(D_{\pi(i)})$ to $F(D'_i)$, $i = 1, \ldots, k$, do step 3 and step 4.
3. For $i = 1, 2, \ldots, k$, suppose that $v_{i,1}, \ldots, v_{i,t}$ are the outer vertices of $D_{\pi(i)}$, and that $v'_{i,1}, \ldots, v'_{i,t}$ are the outer vertices of D'_i, such that $\phi_i(v_{i,j}) = v'_{i,j}$, for $j = 1, \ldots, t$. Replace the cutedge-free component $D_{\pi(i)}$ in G_1 by a wheel

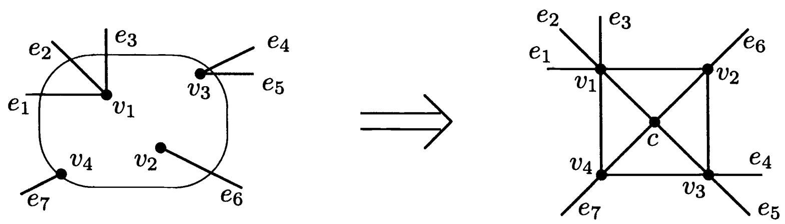


FIG. 3. *Replacing a noncycle cutedge-free component by a wheel.*

$W(c_i; v_{i,1}, \ldots, v_{i,t})$, and replace the cutedge-free component D'_i in G_2 by a wheel $W(c'_i; v'_{i,1}, \ldots, v'_{i,t})$, where c_i and c'_i are new introduced vertices (see Fig. 3 for illustration of this transformation, where $v_1, \ldots, v_4$ are outer vertices of the noncycle cutedge-free component, and e_1, ..., e_7 are cutedges incident on the outer vertices). Moreover, add $2i$ self-loops to each of the centers c_i and c'_i, add one self-loop to each of the vertices $v_{i,1}$ and $v'_{i,1}$, and add two self-loops to each of the vertices $v_{i,2}$ and $v'_{i,2}$, if $v_{i,2}$ and $v'_{i,2}$ exist. Let the resulting graphs be H_1 and H_2, respectively.

4. If the graphs H_1 and H_2 are isomorphic, then the graphs G_1 and G_2 are isomorphic.

We analyze the algorithm. Since the number of noncycle cutedge-free components of G_i, $i = 1, 2$, is less than $3\lceil\Delta\rceil$, then $k < 3\lceil\Delta\rceil$. Therefore, the number of permutations of $(1, \ldots, k)$ is bounded by a constant that is independent of the size of G_1 and G_2. For each permutation π of $(1, \ldots, k)$, the number of different isomorphisms from $D_{\pi(i)}$ to D'_i that are induced by isomorphisms from $F(D_{\pi(i)})$ to $F(D'_i)$ is bounded by $((8\lceil\Delta\rceil)^3)!$, for $i = 1, \ldots, k$, as we have discussed in the analysis of Algorithm 1. Therefore, the number of lists $(\phi_1, \phi_2, \ldots, \phi_k)$, where ϕ_i is an isomorphism from $D_{\pi(i)}$ to D'_i induced by an isomorphism from $F(D_{\pi(i)})$ to $F(D'_i)$, is also bounded by a constant independent of the size of G_1 and G_2. Consequently, the loop of steps 3–4 is executed at most constant many times.

Step 3 can be easily done in linear time. To determine the isomorphism of the graphs H_1 and H_2, note that every cutedge-free component of H_1 and H_2 is planar, since it is either a wheel plus a few self-loops, or a simple cycle. Therefore, the graphs H_1 and H_2 are planar graphs. Using the linear-time algorithm by Hopcroft and Wong [15], we can test the isomorphism of H_1 and H_2 in linear time.

Therefore, Algorithm 2 runs in linear time.

To see the correctness of the algorithm, note that if G_1 and G_2 are isomorphic, then there is an isomorphism ϕ from G_1 to G_2, and a permutation π of $(1, 2, \ldots, k)$ such that, for $i = 1, \ldots, k$, the restriction of ϕ on $F(D_{\pi(i)})$ is an isomorphism ϕ_i from $F(D_{\pi(i)})$ to $F(D'_i)$. The list $(\phi_1, \ldots, \phi_k)$ of these isomorphisms will eventually be found in step 2 of Algorithm 2. Based on this list, we construct a mapping ψ from H_1 to H_2 in the following way: For the wheel $W(c_i; v_{i,1}, \ldots, v_{i,t})$, which corresponds to the cutedge-free component $D_{\pi(i)}$ of G_1, ψ maps c_i to c'_i, and maps $v_{i,j}$ to $v'_{i,j}$, for $j = 1, \ldots, t$, where $W(c'_i; v'_{i,1}, \ldots, v'_{i,t})$ is the wheel in H_2 corresponding to the cutedge-free component D'_i such that $v'_{i,j} = \phi_i(v_{i,j})$, $j = 1, \ldots, t$. For other vertices and edges that do not belong to any noncycle cutedge-free component, ψ is identical with ϕ. It is easy to see that ψ is an isomorphism of H_1 and H_2.

Conversely, suppose that for some permutation π of $(1, \ldots, k)$, and some list of isomorphisms $(\phi_1, \ldots, \phi_k)$, where ϕ_i is an isomorphism from $D_{\pi(i)}$ to D'_i induced by an isomorphism from $F(D_{\pi(i)})$ to $F(D'_i)$, for $i = 1, \ldots, k$, the graphs H_1 and H_2 constructed in Algorithm 2 are isomorphic. Let ψ be an isomorphism of H_1 and H_2. Consider the two wheels $W(c_i; v_{i,1}, \ldots, v_{i,t})$ and $W(c'_i; v'_{i,1}, \ldots, v'_{i,t})$ of H_1 and H_2, which correspond to the two noncycle cutedge-free components $D_{\pi(i)}$ and D'_i, respectively. Since the vertices c_i and c'_i are the only vertices in H_1 and in H_2, respectively, that have $2i$ self-loops, the isomorphism ψ must map c_i to c'_i. Furthermore, the vertices $v_{i,1}$ and $v'_{i,1}$ are the only two vertices that are adjacent to c_i and c'_i in H_1 and in H_2, respectively, and have a single self-loop on them, so ψ must map $v_{i,1}$ to $v'_{i,1}$. Similarly, ψ maps $v_{i,2}$ to $v'_{i,2}$. Now, by the structure of a wheel, ψ must map $v_{i,j}$ to $v'_{i,j}$, for $j = 3, \ldots, t$. Therefore, if we let ϕ be the mapping from G_1 to G_2 that is identical with ψ on the vertices of H_1 that are not the center of the wheels $W(c_i; v_{i,1}, \ldots, v_{i,t})$, $i = 1, \ldots, k$, and is identical with ϕ_i on the vertices in the noncycle cutedge-free component $D_{\pi(i)}$, for $i = 1, \ldots, k$, then it is easy to see that ϕ is an isomorphism from G_1 to G_2.

Acknowledgments. The author thanks Don Friesen, Jonathan Gross, Jan van Leeuwen, Eugene Luks, William Trotter, Chee Yap, and an anonymous referee for many suggestions and stimulating discussions. Their comments led to a substantial improvement in the presentation of the results.

REFERENCES

[1] J. Chen, *A linear time algorithm for isomorphism of graphs of bounded average genus*, Lecture Notes in Comput. Sci., 657 (1993), pp. 103–113.

[2] J. Chen and J. L. Gross, *Limit points for average genus* (I): *3-connected and 2-connected simplicial graphs*, J. Combin. Theory Ser. B, 55 (1992), pp. 83–103.

[3] ———, *Limit points for average genus (II): 2-connected non-simplicial graphs*, J. Combin. Theory Ser. B, 56 (1992), pp. 108–129.

[4] ———, *Kuratowski-type theorems for average genus*, J. Combin. Theory Ser. B, 57 (1993), pp. 100–121.

[5] ———, *No lower limit points for average genus*, in Graph Theory, Combinatorics, Algorithms, and Applications, Y. Alavi et al., eds., Wiley Interscience, New York, 1994, to appear.

[6] I. S. Filotti and J. N. Mayer, *A polynomial-time algorithm for determining the isomorphism of graphs of fixed genus*, in Proc. 12th Annual ACM Symposium on Theory of Comput., Los Angeles, CA, April 1980, pp. 236–243.

[7] M. L. Furst, J. E. Hopcroft, and E. Luks, *A subexponential algorithm for trivalent graph isomorphism*, Tech. report 80–426, Computer Science Department, Cornell University, Ithaca, NY, 1980.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.

[9] J. L. Gross and M. L. Furst, *Hierarchy for imbedding-distribution invariants of a graph*, J. Graph Theory, 11 (1987), pp. 205–220.

[10] J. L. Gross, E. W. Klein, and R. G. Rieper, *On the average genus of graphs*, Graphs Combin., 9 (1993), pp. 153–162.

[11] J. L. Gross, D. P. Robbins, and T. W. Tucker, *Genus distributions for bouquets of circles*, J. Combin. Theory Ser. B, 47 (1989), pp. 292–306.

[12] J. L. Gross and T. W. Tucker, *Topological Graph Theory*, Wiley-Interscience, New York, 1987.

[13] ———, *Stratified graphs*, Discrete Mathematics, to appear.

[14] J. E. Hopcroft and R. E. Tarjan, *Isomorphism of planar graphs*, in Complexity of Computer Computations, R. Miller and J. Thatcher, eds., Plenum, New York, 1972, pp. 131–152.

[15] J. E. Hopcroft and J. K. Wong, *Linear time algorithm for isomorphism of planar graphs*, in Proc. 6th Annual ACM Symposium on Theory of Comput., Seattle, WA, April 1974, pp. 172–184.

[16] E. M. LUKS, *Isomorphism of graphs of bounded valence can be tested in polynomial time*, J. Comput. System Sci., 25 (1982), pp. 42–65.

[17] R. D. RINGEISEN, *Determining all compact orientable* 2*-manifolds upon which* $K_{m,n}$ *has* 2*-cell embeddings*, J. Combin. Theory, 12 (1972), pp. 101–104.

[18] W. T. TUTTE, *Connectivity in Graphs*, University of Toronto Press, Tornto, 1966.

[19] ———, *Graph Theory*, Addison–Wesley, Reading, MA, 1984.

[20] A. T. WHITE, *Graphs, Groups, and Surfaces*, North-Holland, Amsterdam, 1984.

[21] N. H. XUONG, *How to determine the maximum genus of a graph*, J. Combin. Theory Ser. B, 26 (1979), pp. 212–225.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 632–646, November 1994

011

OPTIMAL PARALLEL ALGORITHMS FOR STRAIGHT-LINE GRID EMBEDDINGS OF PLANAR GRAPHS*

MING-YANG KAO†, MARTIN FÜRER‡, XIN HE§, AND BALAJI RAGHAVACHARI¶

Abstract. A straight-line grid embedding of a planar graph is a drawing of the graph on a plane where the vertices are located at grid points and the edges are represented by nonintersecting segments of straight lines joining their incident vertices. Given an n-vertex embedded planar graph with $n \geq 3$, a straight-line embedding on a grid of size $(n-2) \times (n-2)$ can be computed deterministically in $O(\log n \log\log n)$ time with $n/\log n \log\log n$ processors. If randomization is used, the complexity is improved to $O(\log n)$ expected time with the same optimal linear work. These algorithms run on a parallel random access machine that allows concurrent reads and concurrent writes of the shared memory and permits an arbitrary processor to succeed in case of a write conflict.

Key words. planar graphs, straight-line grid embeddings, parallel algorithms

AMS subject classifications. 05C10, 05C85, 68Q22, 68Q25, 68R10

1. Introduction.

A *planar* graph is one that can be drawn on a plane where the vertices are located at points, and the edges are indicated by nonintersecting curves joining their endpoints [3], [4], [5], [12], [16], [21], [32]. A *straight-line grid embedding* of a planar graph is a drawing where the vertices are located at grid points, and each edge is represented by a segment of a straight line. Such embeddings on reasonably small grids are very useful in visualizing planar graphs on graphic screens and have wide applications in CAD/CAM and computer graphics [11], [34].

Wagner [35], Fáry [13], and Stein [31] showed that every planar graph has a straight-line embedding. Since then, many embedding algorithms have been reported [6], [11], [26], [33]. The earlier ones all suffer from two serious drawbacks that render them useless in practice. First, they require high-precision real arithmetic relative to the size of the input graph, and therefore cannot be used even for a graph of moderate size. Second, in the drawings produced by them, the ratio of the smallest distance to the largest distance between vertices is often so small that it is extremely difficult to view those drawings on graphic screens.

In view of these drawbacks, Rosenstiehl and Tarjan [27] posed the problem of computing a straight-line embedding on a grid of polynomial size. Schnyder [29] proved that an n-vertex planar graph with $n \geq 3$ has an embedding on a grid of size $(2n-4) \times (2n-4)$. Independently, de Fraysseix, Pach, and Pollack [10] showed that a straight-line embedding on a grid of size $(2n-4) \times (n-2)$ can be computed in $O(n \log n)$ sequential time. The running time of their algorithm was improved by Chrobak and Payne [7] to optimal $O(n)$ while the grid size remained the same. Schnyder [28] further proved the existence of an embedding on a smaller grid of size

* Received by the editors October 23, 1991; accepted for publication (in revised form) August 6, 1993.

† Department of Computer Science, Duke University, Durham, North Carolina 27708. The research of this author was supported in part by National Science Foundation grant CCR-9101385.

‡ Department of Computer Science, Pennsylvania State University, University Park, Pennsylvania 16802. The research of this author was supported in part by National Science Foundation grants CCR-8805978 and CCR-9218309.

§ Department of Computer Science, State University of New York, Buffalo, New York 14260. The research of this author was supported in part by National Science Foundation grants CCR-9011214 and CCR-9205982.

¶ Department of Computer Science, Pennsylvania State University, University Park, Pennsylvania 16802.

$(n-2) \times (n-2)$ and gave an algorithm [30] to compute such an embedding in optimal $O(n)$ sequential time.

In an earlier paper [14], we showed that a straight-line embedding on a grid of size $(n-2) \times (n-2)$ can be computed in $O(\log n \log\log n)$ time with $n/\log n$ processors on a parallel random access machine (PRAM). If randomization is used, the complexity is improved to $O(\log n)$ expected time with the same nonoptimal $O(n \log\log n)$ work.

This paper presents two optimal linear work algorithms for computing straight-line embeddings on grids of the same small size. We assume that a combinatorial embedding of the input graph is part of the input. (A *combinatorial embedding* [3], [4], [5], [12], [16], [21], [32] is a specification of the exterior boundary and the cyclic order of the edges incident with each vertex in a drawing of a planar graph.) Our deterministic algorithm runs in $O(\log n \log\log n)$ time with $n/\log n \log\log n$ processors, and our randomized algorithm runs in $O(\log n)$ expected time with $n/\log n$ processors. The model of parallel computation is a parallel random access machine (Arbitrary-CRCW PRAM [18]) that allows concurrent reads and concurrent writes of the shared memory, and permits an arbitrary processor to succeed in case of a write conflict. Currently, the best parallel algorithm for computing a combinatorial embedding of a planar graph runs deterministically in $O(\log n)$ time with $O(n \log\log n)$ work, and if randomization is used, it runs in $O(\log n)$ expected time with $O(n)$ work [25].

Our parallel embedding algorithms build upon the sequential embedding algorithm of Schnyder [30]. His algorithm exploits the elegant notion of a realizer of a planar graph [29]. Intuitively, a realizer is a partition of a triangular embedded planar graph into three well-structured trees. Given a realizer, a straight-line grid embedding can be computed, very surprisingly, simply by counting the numbers of vertices in subtrees and tree paths of the realizer. To find a realizer, suitable edges of the input graph are contracted one at a time until the graph is of constant size and has a realizer which can easily be computed. Then, from that realizer, a realizer of the original input graph can be constructed iteratively by adding back the edges and vertices lost in the previous contractions. This useful technique of sequential edge contraction was developed from the works of Wagner [35], Fáry [13], Stein [31], and Kampen [17].

To compute a straight-line grid embedding in parallel, we first show that once a realizer is obtained, a desired embedding can be found with the optimal linear work by processing the trees of the realizer with classic tree-contraction techniques [1], [2], [8], [19], [23], [24]. Next, to compute a realizer, we parallelize the edge contraction technique mentioned above. Note that individually contractible edges need not be simultaneously contractible. Our parallelization techniques make use of fundamental properties of planar graphs. The properties ensure that a constant fraction of the edges in a planar graph are suitable for simultaneous contractions. By contracting such a large set of edges, the size of the input graph can be reduced by a constant fraction. Proceeding recursively, in $O(\log n)$ iterations of contracting edges, the input graph becomes a constant-size graph and has a trivial realizer. That realizer is then expanded into one for the original input graph in $O(\log n)$ iterations of undoing the previous edge contractions. The $O(\log n)$ bound on the number of iterations is sufficient for obtaining an NC algorithm. However, we wish to find a realizer with the optimal $O(n)$ work. To this end, we develop several subroutines that compute a large set of simultaneously contractible edges with various degrees of efficiency, sometimes in constant time. Using load-balancing techniques, we can organize these subroutines to compute a straight-line embedding on a grid of size $(n-2) \times (n-2)$ within the desired $O(n)$ total work.

The above discussion has highlighted the major ideas used in our parallel embedding algorithms. We describe the algorithms as follows. Section 2 reviews the notion of a realizer of a planar graph. It also shows how to compute a straight-line grid embedding from a given realizer and outlines our parallel algorithms for computing straight-line grid embeddings.

The key contribution of our parallel algorithms is finding a realizer with the optimal $O(n)$ work. Section 3 presents an algorithm that computes a realizer by contracting simultaneously contractible edges. Section 4 shows how to compute a large number of simultaneously contractible edges. Section 5 combines the results of the preceding two sections to compute a realizer with the desired $O(n)$ work. Section 6 concludes this paper by proving that a straight-line embedding on a grid of size $(n-2) \times (n-2)$ can be found, indeed, within the claimed $O(n)$ work bound.

2. Computing a straight-line grid embedding from a realizer. Throughout this paper, let G denote an n-vertex *triangular* embedded planar graph with $n \geq 4$. For simplicity, our discussion of how to compute a desired grid embedding focuses on such G and deals with the general case only in §6.

Let v_1, v_2, v_3 be the three exterior vertices of G in the clockwise order. A *realizer* of G is a partition of the interior edges into three sets T_1, T_2, T_3, together with an orientation of the interior edges such that the following properties hold:

1. For each $i \in \{1,2,3\}$, all interior edges incident to the exterior vertex v_i are in T_i and are directed toward v_i.
2. For each interior vertex u in G, the edges incident with u appear around u clockwise in the following pattern:
   - one edge in T_1 leaves u; a set (maybe empty) of edges in T_3 enters u;
   - one edge in T_2 leaves u; a set (maybe empty) of edges in T_1 enters u;
   - one edge in T_3 leaves u; a set (maybe empty) of edges in T_2 enters u.

An example of a realizer is given in Fig. 1. The next theorem makes the notion of a realizer widely applicable and gives some useful tree structures of a realizer.

THEOREM 2.1 (see Schnyder [30]). *Every triangular embedded planar graph has a realizer. For each $i \in \{1,2,3\}$, the edge subset T_i forms a tree in G that is rooted at v_i with all its edges pointing toward the root and consists of all interior vertices and exactly one exterior vertex v_i.*

A straight-line grid embedding can be computed with the following formulas.

For each $i \in \{1,2,3\}$ and for each vertex u in T_i,

- let $P_i(u)$ be the directed tree path in T_i from u to the root v_i of T_i;
- let $p_i(u)$ be the number of vertices in the path $P_i(u)$.

For each $i \in \{1,2,3\}$,

- for each vertex u in T_i, let $t_i(u)$ be the number of vertices in the subtree of T_i rooted at u, and for each exterior vertex $v_j \neq v_i$, let $t_i(v_j) = 1$;
- for each interior vertex u, let

$$r_i(u) = -t_i(u) + \sum_{x \in P_{i-1}(u)} t_i(x) + \sum_{x \in P_{i+1}(u)} t_i(x),$$

where $P_0(u)$ denotes $P_3(u)$, and $P_4(u)$ denotes $P_1(u)$.

The following theorem is the core of Schnyder's algorithm for straight-line grid embeddings.

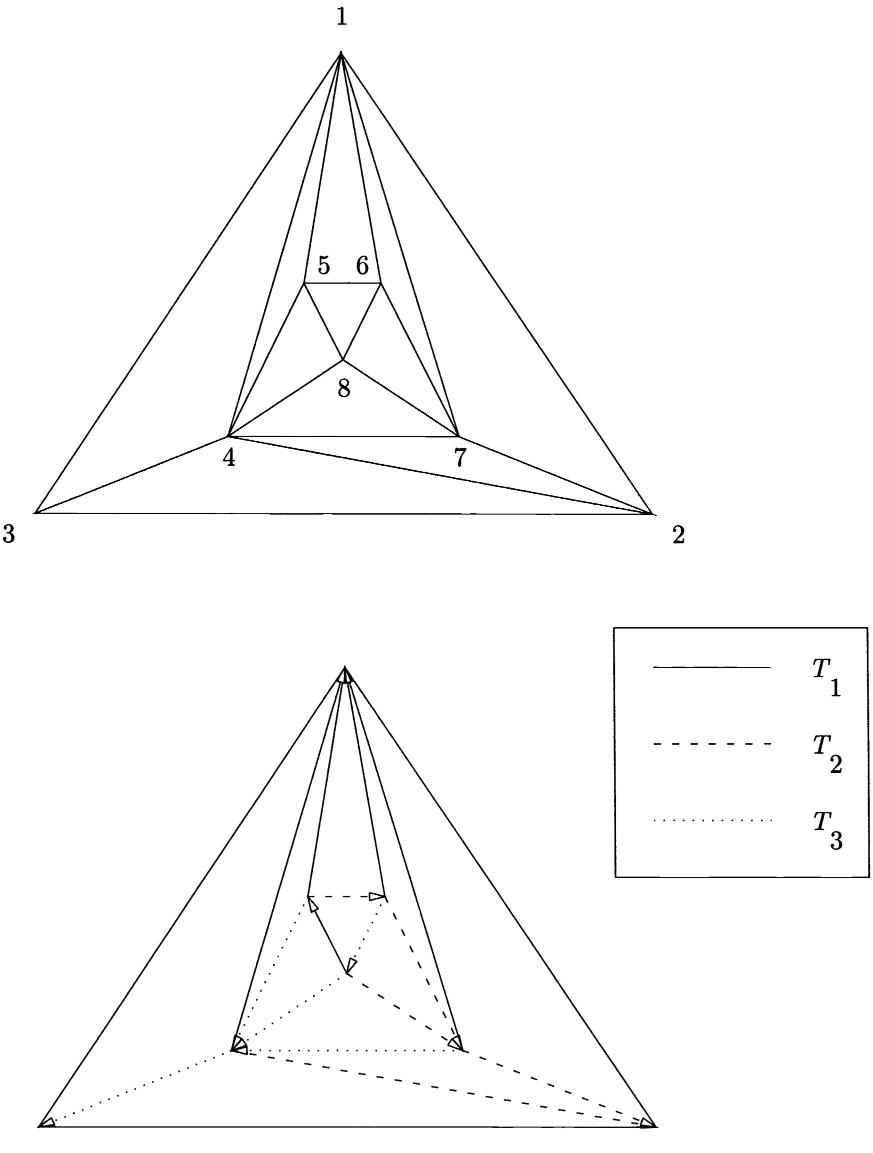


FIG. 1. *Example of a realizer.*

THEOREM 2.2 (see Schnyder [30]). *A straight-line embedding of G on a $(n-2) \times (n-2)$ grid is given by assigning each interior vertex u to the grid point $(r_2(u) - p_1(u), r_1(u) - p_3(u))$ and assigning v_1, v_2, v_3 to $(1, n-2), (n-2, 0), (0, 1)$, respectively.*

Figure 2 shows an example of an embedding given by the above theorem. Figure 3 outlines Schnyder's sequential embedding algorithm based on the theorem. Our parallel algorithms implement this algorithm. It assumes that the combinatorial embedding of G is encoded by a doubly linked circular list of the edges incident with each vertex in the clockwise order. The most difficult part of our parallel algorithms is in showing that at step 1 in Fig. 3, a realizer of G can be computed efficiently. With a realizer found, the integers $p_i(u)$, $t_i(u)$, $r_i(u)$ can be computed in step 2 in $O(\log n)$

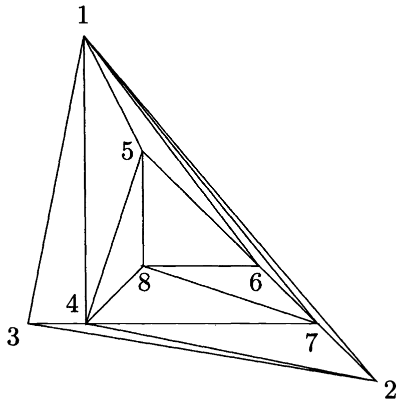


FIG. 2. *Example of a straight-line embedding.*

```
**Input**: $G$.
**Output**: a straight-line embedding of $G$ on a grid of size $(n-2) \times (n-2)$.
**begin**
    1. Find a realizer $T_1, T_2, T_3$ of $G$.
    2. For all interior vertices $u$, compute $p_i(u), t_i(u), r_i(u)$ for $i = 1, 2, 3$.
    3. For all interior vertices $u$, assign $u$ to $(r_2(u) - p_1(u), r_1(u) - p_3(u))$.
    4. Assign $v_1, v_2, v_3$ to $(1, n-2), (n-2, 0), (0, 1)$, respectively.
**end**.
```

FIG. 3. *Computing a straight-line grid embedding.*

time with $n/\log n$ processors by processing T_1, T_2, T_3 with tree contraction [1], [2], [8], [19], [23], [24]. These integers can then be used to compute the coordinates of the vertices in Steps 3 and 4 in $O(1)$ time with n processors. The rest of the paper focuses on finding a realizer of G efficiently in parallel with $O(n)$ work.

3. Computing a realizer by contracting a large set of edges.

In §3.1, we generalize Kampen's notion of a contractible edge [17] and analyze the conditions under which several edges can be contracted simultaneously. In §3.2, we show how to undo these contractions to compute a realizer.

3.1. Two types of simultaneously contractible edge sets.

An edge $e = \{u, v\}$ of G is *contractible* if e is an interior edge and the two endpoints of e have exactly two common neighbors. Let x and y be the two common neighbors of u and v. The contraction of e is performed by merging u and v into a new vertex z. The edges $\{u, x\}$ and $\{v, x\}$ are replaced by a single edge $\{z, x\}$. Similarly $\{u, y\}$ and $\{v, y\}$ are replaced by $\{z, y\}$. All other neighbors of u and v are made adjacent to z. Note that the graph stays triangular after the contraction.

Next, we introduce two types of sets of contractible edges that are suitable for efficient parallel contraction.

An *independent set* of an undirected graph is a set of vertices that are not adjacent to one another. Let S be a set of contractible edges of G. S is *type*-A *compatible* if

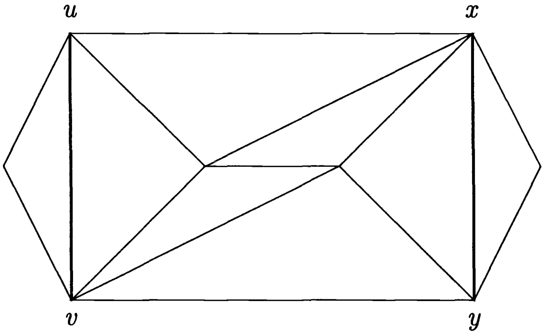


FIG. 4. *Incompatible contractible edges.*

there exists an independent set I of interior vertices of degree three in G such that S consists of one edge incident to each vertex in I. Contracting the edges of S is the same as removing I from G along with the incident edges. The resulting graph is again triangular.

A *quadrangle* of G is a simple cycle of four edges. Contractible edges on opposite sides of a quadrangle may not be contractible simultaneously since contracting one of the edges might make the other edge uncontractible. For example, in Fig. 4 the edges $\{u, v\}$ and $\{x, y\}$ are individually but not simultaneously contractible. In light of this observation, two contractible edges of G are *type*-B *compatible* if they do not share a common vertex and are not on the opposite sides of a quadrangle in G. A set of contractible edges is type-B compatible if its edges are pairwise type-B compatible.

A set S of contractible edges in G is *compatible* if it is type-A compatible or type-B compatible. Let G/S denote the graph obtained by contracting S in G. It is easy to see from the definition of compatibility that G/S is a triangular planar graph and its combinatorial embedding induced from that of G is unique. Also, if S contains m edges, then G/S contains exactly $n - m$ vertices.

LEMMA 3.1. *Let S be a compatible set of contractible edges in G. Given S and G, the unique embedded planar graph G/S can be computed deterministically in $O(1)$ time with n processors.*

Proof. There are two cases: S is type-A or type-B compatible.

Case 1. S is type-A compatible. Let I be the set of interior vertices of degree three in G that induces S. Then the graph G/S can be constructed from G by deleting all vertices in I. Also, each interior face of G/S contains at most one vertex of I. Based on these observations, each vertex of I can be deleted locally in $O(1)$ time by a processor allocated to that vertex. Thus the total complexity is as desired.

Case 2. S is type-B compatible. Let G' be the embedded planar graph constructed from G by topologically shrinking each edge in S into a single vertex. Then G/S can be constructed by deleting redundant multiple edges in G'. Because the edges in S are disjoint, G' and its combinatorial embedding can be computed in $O(1)$ time with n processors. Each multiple edge in G' has exactly two copies, and they are on the boundary of a face of size two. Therefore, the redundant multiple edges in G' can be deleted in $O(1)$ time with n processors, and the total complexity is as stated. □

3.2. Undoing the contractions of compatible sets.

By Theorem 2.1, both G/S and G have realizers. Given a realizer T_1, T_2, T_3 of G/S, we can compute a realizer of G by applying to each edge in S a replacement procedure to be described below.

Because the edges in S are interior edges, the exterior vertices of G are not contracted with one another in G/S. Thus, G and G/S have the same exterior vertices. The realizer T_1, T_2, T_3 and the realizer of G to be computed by the replacement procedure are rooted at the same exterior vertices in the same clockwise order. Hence, for notational simplicity, we use the same notations T_1, T_2, T_3 for the desired realizer of G.

To describe the replacement procedure, additional notations are in order. Let e be an edge in S. Let x and y be the endpoints of e. Let u and w be the two common neighbors of x and y in G. Assume that x, u, y, and w are in the clockwise order on the quadrangle formed by them.

Let z be the vertex in G/S that is contracted from x and y. Let u' and w' be the vertices in G/S that are contracted from u and w or are u and w themselves. Note that u' and w' are different vertices in G/S because by the compatibility of S, there is no edge in S between u and w.

Let d_1 be the directed edge in $T_1 \cup T_2 \cup T_3$ between u' and z. Let d_2 be the directed edge in $T_1 \cup T_2 \cup T_3$ between w' and z. Let E_x be the set of edges in G/S that are different from d_1 and d_2, and are originally incident with x in G.

To obtain a realizer for G, the edges d_1 and d_2 in T_1, T_2, T_3 need to be replaced with the five edges in G between x, y, u, and w. This replacement is done for the d_1 and d_2 of each e in S. All other edges keep their directions and remain in the same trees T_i.

The five edges in G between x, u, y, and w need to be assigned appropriate directions and trees T_i. The assignments are determined by the following four factors:

- which of T_1, T_2, and T_3 the edges d_1 and d_2 belong to;
- the directions of d_1 and d_2;
- whether x is an exterior vertex in G;
- whether E_x contains at least one outgoing edge from z.

The appropriate directions and trees T_i for the five edges between x, u, y, and w in G are detailed in the case analysis in Fig. 5. The analysis covers all possible cases up to the symmetries of rotating T_1, T_2, T_3 and rotating x, u, y, w.

LEMMA 3.2. *The replacement rules in Fig.* 5 *correctly produce a realizer of* G.

Proof. The correctness of the replacement rules is by straightforward verification based on the following observations concerning cyclic edge patterns of a realizer.

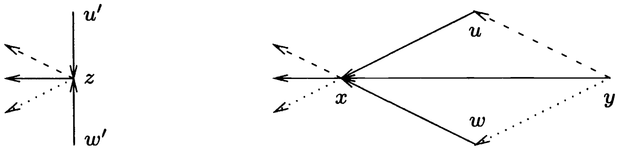
First, in each of the cases in Fig. 5, d_1 is replaced by two directed edges between u and x, y. The crucial fact is that the subpattern formed by d_1 in the cyclic edge pattern around u' is preserved by the subpattern formed by the two substitute edges around u. Therefore, from the standpoint of u, the replacement for d_1 preserves the cyclic edge pattern around u' that is required for a realizer. The same preservation of a cyclic edge pattern holds from the standpoint of w.

Similarly, in each of the cases in Fig. 5, z is split into x and y. The subpattern formed by d_2, E_x, d_1 in the cyclic edge pattern around z is preserved by the three edges between y and w, x, u around y. Thus, from the standpoint of y, the splitting of z into x and y maintains the cyclic edge pattern around z that is required for a realizer. The same pattern preservation holds from the standpoint of x.

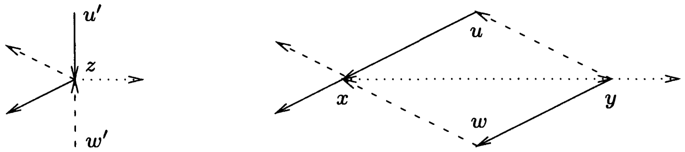
These pattern preservations together with the cyclic edge patterns of T_1, T_2, T_3, ensure that the replacement rules in Fig. 5 indeed produce a realizer for G. □

LEMMA 3.3. *Let* S *be a compatible set of contractible edges in* S. *If* G, S, G/S, *and a realizer of* G/S *are given, then a realizer of* G *can be computed deterministically*

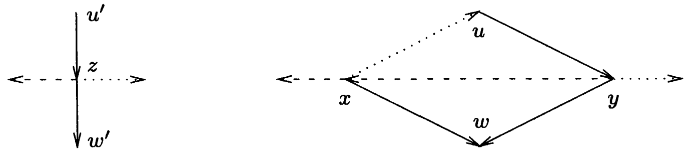
1. If $d_1 = u' \to z \in T_1$ and $d_2 = w' \to z \in T_1$, and if either x is an exterior vertex of G or E_x contains at least one outgoing edge from z, then replace d_1 and d_2 with the following edges: $y \to x \in T_1, u \to x \in T_1, y \to u \in T_2, y \to w \in T_3, w \to x \in T_1$.



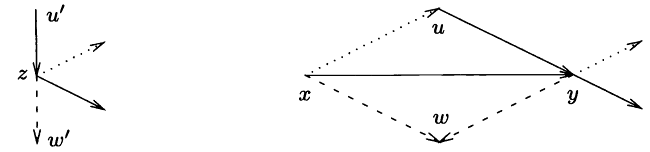
2. If $d_1 = u' \to z \in T_1$ and $d_2 = w' \to z \in T_2$, then replace d_1 and d_2 with the following edges: $x \to y \in T_3, u \to x \in T_1, y \to u \in T_2, y \to w \in T_1, w \to x \in T_2$.



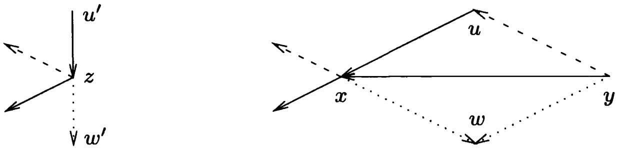
3. If $d_1 = u' \to z \in T_1$ and $d_2 = z \to w' \in T_1$, then replace d_1 and d_2 with the following edges: $y \to x \in T_2, x \to u \in T_3, u \to y \in T_1, y \to w \in T_1, x \to w \in T_1$.



4. If $d_1 = u' \to z \in T_1$ and $d_2 = z \to w' \in T_2$, then replace d_1 and d_2 with the following edges: $x \to y \in T_1, x \to u \in T_3, u \to y \in T_1, y \to w \in T_2, x \to w \in T_2$.



5. If $d_1 = u' \to z \in T_1$ and $d_2 = z \to w' \in T_3$, then replace d_1 and d_2 with the following edges: $y \to x \in T_1, u \to x \in T_1, y \to u \in T_2, y \to w \in T_3, x \to w \in T_3$.



6. If $d_1 = z \to u' \in T_1$ and $d_2 = z \to w' \in T_2$, then replace d_1 and d_2 with the following edges: $y \to x \in T_3, x \to u \in T_1, y \to u \in T_1, y \to w \in T_2, x \to w \in T_2$.

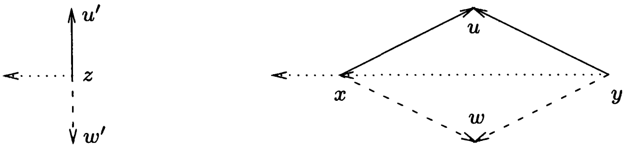


FIG. 5. *Replacement rules for undoing contractions.*

in $O(1)$ time with n processors.

Proof. In light of Lemma 3.2, we use the replacement rules in Fig. 5 to produce a realizer for G. The four determining factors for each edge in S can be computed within the desired complexity. Then, because there is only a constant number of replacement rules, the right rule for each edge can also be determined within the desired complexity. Next, the replacement at each edge is simple enough to be performed locally at that edge within the desired complexity. □

4. Computing a large set of simultaneously contractible edges.

We have shown that a compatible set can be contracted and expanded to compute a realizer in parallel. It remains for us to show how to find such sets with a constant fraction of the edges in G. Our approach is to compute a large independent set of G that has useful neighborhood properties. We then use this independent set to compute two compatible sets that together consist of a large number of edges in G.

4.1. Finding a large independent set.

Let J be an independent set of G. For $x \in J$, a neighbor y of x is a *bad* neighbor if y is also a neighbor of another vertex in J, otherwise y is a *good* neighbor. Let $G = (V, E)$. For $x \in V$, let $\deg(x)$ denote the degree of x in G. Let $V_i = \{x \in V | \deg(x) = i\}$ and $V_{[i,j]} = \{x \in V | i \leq \deg(x) \leq j\}$. The vertices in $V_{[0,29]}$ are the *light* vertices of G; the vertices in $V_{[30,\infty]}$ are the *heavy* vertices.

In our embedding algorithms, we wish to find a large independent set of interior vertices such that for every vertex in that set, all its neighbors are good. Unfortunately, this is not always possible. Instead, we adopt the following weaker approach. The independent set J is *sparse* in G if the following conditions hold for all $x \in J$:

- $\deg(x) \leq 6$;
- all light neighbors of x are good;
- if $\deg(x) = 4$, then at least two neighbors of x are light;
- if $\deg(x) = 5$, then at least four neighbors of x are light;
- if $\deg(x) = 6$, then all neighbors of x are light.

Below we describe a construction of nine steps that finds a large sparse independent set for G. It actually works for any planar graph with a minimum degree of at least three and has been used for five-coloring a planar graph [15].

1. Let $U_4 = \{x \in V_4 \mid x \text{ has at least two light neighbors}\}$.
2. Let $U_5 = \{x \in V_5 \mid x \text{ has at least four light neighbors}\}$.
3. Let $U_6 = \{x \in V_6 \mid \text{all neighbors of } x \text{ are light}\}$.
4. Let $U = V_3 \cup U_4 \cup U_5 \cup U_6$.
5. Let $G_1 = (V_{[3,29]}, E_1)$ be the subgraph of G induced by $V_{[3,29]}$.
6. Let $G_1^s = (V_{[3,29]}, E_1^s)$ be the graph such that $(x, z) \in E_1^s$ if and only if there is a vertex y such that both $(x, y), (y, z) \in E_1$.
7. Let $G_2 = (V_{[3,29]}, E_1 \cup E_1^s)$.
8. Let G_3 be the subgraph of G_2 induced by the set U.
9. Let J be an independent set of G_3.

LEMMA 4.1. *Every independent set J of G_3 is a sparse independent set of G.*

Proof. First of all, because $U \subseteq V_{[3,29]}$, Steps 5, 7, and 8 ensure that J is an independent set of G. By steps 8 and 9, the vertices in J are of degree six at most. By steps 6, 7, and 8, all light neighbors must be good. By step 1, if a vertex $x \in J$ is of degree four in G, then x has at least two light neighbors in G. By step 2, if a vertex $x \in J$ is of degree five in G, then x has at least four light neighbors. By step 3, if a vertex $x \in J$ is of degree six in G, then all neighbors of x are light. □

LEMMA 4.2. *The set U defined above contains at least $\lceil n/48 \rceil$ vertices.*

Proof. This proof relies on the fact that the minimum degree of G is at least three. Let $n_i = |V_i|$, $n_{[i,j]} = |V_{[i,j]}|$, and $p_i = |U_i|$. We need to show

$$n_3 + p_4 + p_5 + p_6 \geq \frac{n}{48}.$$

If $n_3 \geq \frac{n}{48}$, we are done. Thus we assume that $n_3 < \frac{n}{48}$. We first show the following inequality by contradiction:

$$(1) \qquad n_{[30,\infty]} < \frac{n}{8}.$$

If $n_{[30,\infty]} \geq n/8$, then the sum of the degrees of all the vertices is at least

$$30 \cdot \frac{n}{8} + 3 \cdot \frac{7n}{8} > 6n.$$

This is a contradiction because by Euler's formula any planar graph has at most $3n-6$ edges; hence the sum of the degrees of the vertices cannot exceed $6n - 12$.

By counting the edges between $V_{[30,\infty]}$ and $(V_4 - U_4) \cup (V_5 - U_5) \cup (V_6 - U_6)$,

$$3(n_4 - p_4) + 2(n_5 - p_5) + (n_6 - p_6) \leq \sum_{x \in V_{[30,\infty]}} \deg(x).$$

Thus,

$$n_3 + 3n_4 + 2n_5 + n_6 - (n_3 + 3p_4 + 2p_5 + p_6) \leq \sum_{x \in V_{[30,\infty]}} \deg(x).$$

Therefore,

$$n_3 + 3n_4 + 2n_5 + n_6 + \sum_{x \in V_{[3,6]}} \deg(x) + \sum_{x \in V_{[7,29]}} \deg(x) - (n_3 + 3p_4 + 2p_5 + p_6)$$
$$\leq \sum_{x \in V} \deg(x) = 2|E| \leq 6(n_3 + n_4 + n_5 + n_6 + n_{[7,29]} + n_{[30,\infty]}).$$

Simplifying this inequality yields

$$n_3 + n_4 + n_5 + n_6 + n_{[7,29]} - (n_3 + 3p_4 + 2p_5 + p_6) - 3n_3 \leq 6n_{[30,\infty]}.$$

Hence,

$$(2) \qquad n_3 + 3p_4 + 2p_5 + p_6 \geq n - 7n_{[30,\infty]} - 3n_3.$$

By (1) and (2) and the assumption that $n_3 < \frac{n}{48}$,

$$n_3 + p_4 + p_5 + p_6 \geq \frac{1}{3}(n_3 + 3p_4 + 2p_5 + p_6) \geq \frac{1}{3}\left(n - \frac{7}{8}n - \frac{3n}{48}\right) = \frac{n}{48}. \qquad \square$$

For a constant $c > 0$, an independent set of an m-vertex graph is *c-large* if it contains at least $c \cdot m$ vertices.

LEMMA 4.3. *Given a constant $c > 0$, every c-large independent set J of G_3 is a $\frac{c}{48}$-large sparse independent set of G.*

Proof. The sparse independence of J has been shown in Lemma 4.1. The lower bound on the cardinality of J follows directly from Lemma 4.2. □

LEMMA 4.4.

(1) *A sparse independent set of $\Omega(n)$ interior vertices in G can be computed deterministically in $O(\log n)$ time with $n/\log n$ processors.*

(2) *A sparse independent set of $\Omega(n)$ interior vertices in G can be computed deterministically in $O(\log^* n)$ time with n processors.*

(3) *A sparse independent set of interior vertices in G can be computed in $O(1)$ expected time with n processors on a probabilistic* PRAM. *The independent set found has $\Omega(n)$ vertices with probability at least a positive constant.*

Proof. We employ the nine-step procedure described earlier to compute a desired independent set for G. Because the degrees of the vertices of $V_{[3,29]}$ are bounded by a constant in G, the graph G_3 can be constructed in $O(\log n)$ time with $n/\log n$ processors, or in $O(1)$ time with n processors.

Next we compute a large J with three methods. These methods all rely on the fact that the maximum degree of G_3 is bounded by a constant. The first method decomposes G_3 into a constant number of trees and computes J by two-coloring these trees in $O(\log n)$ time with $n/\log n$ processors with tree contraction [1], [2], [8], [19], [23], [24]. The second method computes J in $O(\log^* n)$ time with n processors with parallel symmetry-breaking techniques [15]. The third method computes J with one iteration of Luby's Monte Carlo algorithm for finding a maximal independent set [22]. One iteration of Luby's algorithm runs in $O(1)$ time with $|G_3|$ processors and finds a J of $\Omega(|G_3|)$ vertices with probability at least a positive constant.

By Lemma 4.3, J is also a large sparse independent set of G. If J contains an exterior vertex, we simply delete that vertex, in $O(\log n)$ time with $n/\log n$ processors or in $O(1)$ time with n processors, to get a large sparse independent set of interior vertices of G. Therefore a sparse independent set of $\Omega(n)$ interior vertices in G can be computed within the stated complexity bounds. □

4.2. Computing contractible edges from an independent set. Let I be a large sparse independent set of interior vertices of G. Below, we will show how to compute from I a type-A compatible set S_a and a type-B compatible set S_b of G. We will also prove that S_a and S_b hold two useful conditions: (1) no edge of S_b disappears in G/S_a; and (2) S_b remains a type-B compatible set in G/S_a. We call such S_a and S_b an AB-*pair* of G, and will contract first S_a and then S_b to compute a realizer of G. The two conditions ensure that S_b fits for simultaneous contractions after S_a is contracted. Let n be the number of vertices in G. Let n_a and n_b be the numbers of edges in S_a and S_b, respectively. By the first condition, there are $n - n_a - n_b$ vertices in the new graph resulting from contracting S_a and S_b. We will show that the AB-pair computed from I contains a large number of edges. Thus, the contractions reduce the size of G by a constant fraction.

Given I, the set S_a is computed as follows: for each vertex in I that has degree three, find an edge in G incident with that vertex and include that edge in S_a.

LEMMA 4.5. *S_a is a type-*A *compatible set of G. Given I and G, it can be computed deterministically in $O(1)$ time with n processors.*

Proof. This lemma follows from the fact that every edge incident with an interior vertex of degree three in G is contractible. □

Next we show how to compute S_b. For brevity, a neighbor y of an interior vertex x in G is contractible if the edge $\{x, y\}$ is a contractible one in G.

LEMMA 4.6 (see Kampen [17]). *In a triangular embedded planar graph, every interior vertex has at least two contractible neighbors.*

Given I, the set S_b is constructed by choosing an edge in G incident with each vertex of I with degree four, five, or six as follows.

Case 1. $\deg(x) = 4$. If x has a light contractible neighbor y, then add the edge $\{x, y\}$ to S_b. If x has no light contractible neighbor, then it has two light neighbors p and r, and the other two neighbors q and s are contractible. As p is not contractible, it shares three neighbors with x. These must be q, r, and s. Likewise r shares the neighbors p, q, and s with x. Hence, either q or s is in the interior of the triangle C formed by p, r, and x. If q is in the interior of C, then add the contractible edge $\{x, q\}$ to S_b. If s is in the interior of C, then add the contractible edge $\{x, s\}$ to S_b.

Case 2. $\deg(x) = 5$ or 6. Choose a light contractible neighbor y of x and add the edge $\{x, y\}$ to S_b. The existence of such y follows from Lemma 4.6 and the definition of a sparse independent set.

LEMMA 4.7. *S_b is a type-*B *compatible set of G.*

Proof. First, every edge in S_b is contractible because the construction of S_b always selects a contractible neighbor for each given x. To show S_b is type-B compatible, consider any two vertices x_1 and x_2 in I with degree four, five, or six. Let $e_1 = \{x_1, y_1\}$ and $e_2 = \{x_2, y_2\}$ be the two edges in S_b that are chosen for x_1 and x_2, respectively. If y_1 is a good neighbor of x_1, then e_1 and e_2 cannot share a common vertex. Furthermore, if y_1 is a good neighbor of x_1, then by the independence of I, e_1 and e_2 cannot form opposite sides of a quadrangle in G. Symmetrically, if y_2 is a good neighbor of x_2, then the same claims are true.

Based on the above discussion, below we assume that y_1 is a bad neighbor of x_1 and that y_2 is a bad neighbor of x_2. Then, $\deg(x_1) = \deg(x_2) = 4$ by the construction of S_b. Let p_1, y_1, r_1, s_1 be the four neighbors of x_1 in the clockwise order. Let p_2, y_2, r_2, s_2 be the four neighbors of x_2 in the clockwise order. Since y_1 is a bad neighbor of x_1, y_1 cannot be light by the definition of a sparse independent set. Then, because y_1 is selected for x_1, the following separation properties must hold: (1) x_1, p_1, r_1 form a triangle C_1; (2) y_1 is in the interior of C_1; and (3) s_1 is in the exterior of C_1. Similarly, since y_2 is a bad neighbor of x_2 and is selected for x_2, the following separation properties must hold: (4) x_2, p_2, r_2 form a triangle C_2; (5) y_2 is in the interior of C_2; and (6) s_2 is in the exterior of C_2. By planarity and these separation properties, e_1 and e_2 cannot share an endpoint and cannot form opposite sides of a quadrangle. □

LEMMA 4.8. *Let n denote the number of vertices in G. Given I and G, the set S_b can be found in $O(1)$ time with n processors.*

Proof. For each $x \in I$ of degree four, five, or six, we first determine all light neighbors of x. With all light neighbors of x determined, we then determine which of them are contractible. Because the degrees of x and its light neighbors are all bounded by a constant, these two steps take $O(1)$ time with n processors. Hence S_b can be computed in the desired complexity. □

LEMMA 4.9. *S_a and S_b form an* AB*-pair of G with $|S_a| + |S_b| = |I|$.*

Proof. The AB-pair property follows from the fact that G/S_a can be obtained by deleting from G the vertices of I that have degree three in G. The size property follows from the fact that S_a and S_b together consist of an edge incident with each vertex in I. □

LEMMA 4.10.

(1) *An* AB*-pair of G with $\Omega(n)$ edges can be deterministically computed and con-*

tracted in $O(\log n)$ time with $n/\log n$ processors.

(2) *An* AB*-pair of G with $\Omega(n)$ edges can be deterministically computed and contracted in $O(\log^* n)$ time with n processors.*

(3) *An* AB*-pair of G can be computed and contracted in $O(1)$ expected time with n processors on a probabilistic* PRAM. *The* AB*-pair found has $\Omega(n)$ edges with probability at least a positive constant.*

Proof. This lemma follows directly from Lemmas 4.4, 4.5, 4.8, and 4.9. □

5. Parallel complexities for computing realizers.

Here we give deterministic and randomized complexities for computing a realizer in parallel.

5.1. Deterministic parallel complexity for computing realizers.

THEOREM 5.1. *A realizer of an n-vertex triangular embedded planar graph can be computed deterministically in $O(\log n \log\log n)$ time with $n/\log n \log\log n$ processors.*

To compute a realizer of the input graph G, we first contract G in three stages:

```
1. repeat
       Distribute the vertices and edges of $G$ evenly over the given processors.
       Contract $G$ by means of Lemma 4.10(1).
   until $G$ has at most $n/\log\log n$ vertices.
2. Distribute the vertices and edges of $G$ evenly over the given processors.
   repeat
       Contract $G$ by means of Lemma 4.10(1).
   until $G$ has at most $n/\log n \log\log n$ vertices.
3. Distribute the vertices and edges of $G$ evenly over the given processors.
   repeat
       Contract $G$ by means of Lemma 4.10(2).
   until $G$ is a triangle with a single interior vertex.
```

The processor allocation steps are done with an algorithm [20] that computes the prefix sums of an $O(n)$-cell array in $O(\log n)$ time with $n/\log n$ processors.

After the third contraction stage, G has a trivial realizer with each T_i consisting of one edge. We expand this realizer into one for the original G by undoing the contractions by means of Lemma 3.3. No work is needed to redistribute the processors because the processor assignment at every contraction step can be recorded.

The correctness of the above algorithm follows from Lemmas 4.10(1), 4.10(2), and 3.3. The time complexity follows from the next lemma.

LEMMA 5.2. *The contraction stages take $O(\log n \log\log n)$, $O(\log n \log\log n)$, and $O(\log n \log^* n)$ time, respectively, and they can be undone in $O(\log n \log\log n)$, $O(\log n \log\log n)$, and $O(\log n)$ time, respectively.*

Proof. The proofs for performing the contractions and undoing them are similar; only that for performing the contractions is given below. Let n_i be the number of vertices in G at the start of the ith iteration of the repeat loops.

Stage 1. In the ith iteration, $|G| = O(n_i)$. Thus, the load-balancing step, the contraction step, and the until step take $O(\log n_i)$ time with $n_i/\log n_i$ processors. Because $n_i \geq n/\log\log n$, the number of given processors is smaller than $n_i/\log n_i$. Thus, the running time of this iteration is $O(\frac{n_i}{n} \log n \log\log n)$ with the given processors. Because n_i decreases at least geometrically, the running time of an iteration also decreases at least geometrically and the lemma for stage 1 follows.

Stage 2. Because $|G| = O(n/\log\log n)$ and $n/\log n \log\log n$ processors are given, the load-balancing step takes $O(\log n)$ time. In the ith iteration, the contraction step and the until step take $O(\log n_i)$ time with $n_i/\log n_i$ processors. Because $n_i \leq n/\log\log n$, $n_i/\log n_i = O(n/\log n \log\log n)$ and each iteration takes $O(\log n_i)$ time with the given processors. Because each iteration reduces n_i by at least a constant fraction, $O(\log\log n)$ iterations suffice and the lemma for stage 2 follows.

Stage 3. Because $|G| = O(n/\log n \log\log n)$ and there are $n/\log n \log\log n$ processors, the load-balancing step takes $O(\log n)$ time. In the ith iteration, the contraction step takes $O(\log^* n_i)$ time with n_i processors. The until step takes $O(1)$ time with one processor. Because $n_i \leq n/\log n \log\log n$, each iteration takes $O(\log^* n)$ time with the given processors. Since each iteration reduces n_i by at least a constant fraction, $O(\log n)$ iterations are needed and the lemma for stage 3 follows. □

5.2. Randomized parallel complexity for computing realizers.

THEOREM 5.3. *Given a triangular embedded planar graph with n vertices, a realizer can be computed in $O(\log n)$ expected time with $n/\log n$ processors on a probabilistic* PRAM.

Proof. Our randomized algorithm is identical to the deterministic one given in the proof of Theorem 5.1 with only two differences. First, the three contraction stages contract the input graph by means of Lemma 4.10(3) until it has at most $n/\log\log n$ vertices, at most $n/\log n$ vertices, and only one interior vertex, respectively. Second, the processor allocation steps of these three stages and the until steps of the first two stages are performed with an algorithm [9] that computes the prefix sums of an n-cell array in $O(\log n/\log\log n)$ time and $O(n)$ work. The analysis of this algorithm is similar to that of the deterministic one. □

6. Conclusions.

We are now ready to give the main results of this paper.

THEOREM 6.1. *Given an n-vertex embedded planar graph with $n \geq 3$, a straight-line embedding on a grid of size $(n-2) \times (n-2)$ can be computed deterministically in $O(\log n \log\log n)$ time with $n/\log n \log\log n$ processors. Alternately, such an embedding can be computed in $O(\log n)$ expected time with $n/\log n$ processors on a probabilistic* PRAM.

Proof. The case $n = 3$ is trivial. So we assume $n \geq 4$ and compute an embedding in three stages. First, triangulate the input graph. Next, compute an embedding of the triangulated graph by means of Theorems 5.1 and 5.3 and the discussion at the end of §2. Last, delete the added edges to yield a desired embedding for the original graph. All three stages can be done with the desired complexity. □

Acknowledgments. We would like to thank Deganit Armon, Hillel Gazit, Samir Khuller, and Walter Schnyder for extremely helpful discussions.

REFERENCES

[1] K. ABRAHAMSON, N. DADOUN, D. G. KIRKPATRICK, AND T. PRZYTYCKA, *A simple tree contraction algorithm*, J. Algorithms, 10 (1989), pp. 287–302.

[2] R. J. ANDERSON AND G. L. MILLER, *Deterministic parallel list ranking*, Algorithmica, 6 (1991), pp. 859–868.

[3] C. BERGE, *Graphs*, 2nd rev. ed., North–Holland, New York, 1985.

[4] B. BOLLOBÁS, *Graph Theory*, Springer-Verlag, New York, 1979.

[5] J. BONDY AND U. MURTY, *Graph Theory with Applications*, North–Holland, Amsterdam, 1976.

[6] N. CHIBA, T. YAMANOUCHI, AND T. NISHIZEKI, *Linear algorithms for convex drawings of planar graphs*, in Progress in Graph Theory, Academic Press, Orlando, FL, 1984, pp. 153–173.

[7] M. CHROBAK AND T. H. PAYNE, *A linear time algorithm for drawing planar graphs on a grid*, Tech. report UCR-CS-90-2, Department of Mathematics and Computer Science, University of California at Riverside, Riverside, CA, 1990.

[8] R. COLE AND U. VISHKIN, *The accelerated centroid decomposition technique for optimal tree evaluation in logarithmic time*, Algorithmica, 3 (1988), pp. 329–346.

[9] ———, *Faster optimal prefix sums and list ranking*, Inform. and Comput., 81 (1989), pp. 334–352.

[10] H. DE FRAYSSEIX, J. PACH, AND R. POLLACK, *Small sets supporting Fáry embeddings of planar graphs*, in Proc. 20th Ann. ACM Sympos. on Theory of Computing, Chicago, IL, 1988, pp. 426–433.

[11] P. EADES AND R. TAMASSIA, *Algorithms for drawing graphs: An annotated bibliography*, Tech. report CS-89-09, Department of Computer Science, Brown University, Providence, RI, 1989.

[12] S. EVEN, *Graph Algorithms*, Computer Science Press, Rockville, MD, 1979.

[13] I. FÁRY, *On straight line representation of planar graphs*, Acta Sci. Math. (Szeged), 11 (1948), pp. 229–233.

[14] M. FÜRER, X. HE, M. Y. KAO, AND B. RAGHAVACHARI, *$O(n \log\log n)$-work parallel algorithms for straight-line grid embeddings of planar graphs*, in Proc. 4th Ann. ACM Sympos. on Parallel Algorithms and Architectures, San Diego, CA, 1992, pp. 410–419.

[15] A. GOLDBERG, S. PLOTKIN, AND G. SHANNON, *Parallel symmetry-breaking in sparse graphs*, SIAM J. Discrete Math., 1 (1988), pp. 434–446.

[16] F. HARARY, *Graph Theory*, Addison–Wesley, Reading, MA, 1969.

[17] G. R. KAMPEN, *Orienting planar graphs*, Discrete Math., 14 (1976), pp. 337–341.

[18] R. KARP AND V. RAMACHANDRAN, *A survey of parallel algorithms for shared-memory machines*, in Handbook of Theoretical Computer Science: Algorithms and Complexity, Vol. A, J. van Leeuwen, ed., Elsevier, New York, 1990, pp. 869–941.

[19] S. R. KOSARAJU AND A. L. DELCHER, *Optimal parallel evaluation of tree-structured computations by raking*, in Lecture Notes in Computer Science 319: VLSI Algorithms and Architectures, the 3rd Aegean Workshop on Computing, J. H. Reif, ed., Springer-Verlag, Berlin, Heidelberg, 1988, pp. 101–110.

[20] R. E. LADNER AND M. J. FISCHER, *Parallel prefix computation*, J. Assoc. Comput. Math., 27 (1980), pp. 831–838.

[21] L. LOVÁSZ, *An Algorithmic Theory of Numbers, Graphs and Convexity*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1986.

[22] M. LUBY, *A simple parallel algorithm for the maximal independent set problem*, SIAM J. Comput., 15 (1986), pp. 1036–1053.

[23] G. L. MILLER AND J. H. REIF, *Parallel tree contraction and its application*, in Proc. 26th Ann. IEEE Sympos. on Foundations of Computer Science, Portland, OR, 1985, pp. 478–489.

[24] ———, *Parallel tree contraction, part 1: Fundamentals*, in Advances in Computing Research: Randomness and Computation, Vol. 5, S. Micali, ed., JAI Press, Greenwich, CT, 1989, pp. 47–72.

[25] V. RAMACHANDRAN AND J. H. REIF, *An optimal parallel algorithm for graph planarity*, in Proc. 30th Ann. IEEE Sympos. on Foundations of Computer Science, Research Triangle Park, NC, 1989, pp. 282–287.

[26] R. C. READ, *A new method for drawing a planar graph given the cyclic order of the edges at each vertex*, Congr. Numer., 56 (1987), pp. 31–44.

[27] P. ROSENSTIEHL AND R. TARJAN, *Rectilinear planar layouts and bipolar orientations of planar graphs*, Discrete Comput. Geom., 1 (1986), pp. 343–353.

[28] W. SCHNYDER, *Embedding planar graphs on the grid*, Abstracts Amer. Math. Soc., 9 (1988), p. 268.

[29] ———, *Planar graphs and poset dimension*, Order, 5 (1989), pp. 323–343.

[30] ———, *Embedding planar graphs on the grid*, in Proc. 1st Ann. ACM–SIAM Sympos. on Discrete Algorithms, San Francisco, CA, 1990, pp. 138–148.

[31] S. K. STEIN, *Convex maps*, in Proc. Amer. Math. Soc., Vol. 2, Providence, RI, 1951, pp. 464–466.

[32] W. TUTTE, *Graph Theory*, Encyclopedia Math. Appl., Vol. 21, Addison–Wesley, Reading, MA, 1984.

[33] W. T. TUTTE, *How to draw a graph*, Proc. London Math. Soc., 13 (1963), pp. 743–768.

[34] J. ULLMAN, *Computational Aspects of VLSI*, Computer Science Press, Rockville, MD, 1984.

[35] K. WAGNER, *Bemerkungen zum Vierfarbenproblem*, Jahresber. Deutsch. Math.-Verein., 46 (1936), pp. 26–32.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 647–655, November 1994

TREEWIDTH OF CIRCULAR-ARC GRAPHS*

RAVI SUNDARAM†, KARAN SHER SINGH‡, AND C. PANDU RANGAN§

Abstract. The treewidth of a graph is one of the most important graph-theoretic parameters from the algorithmic point of view. However, computing the treewidth and constructing a corresponding tree-decomposition for a general graph is NP-complete. This paper presents an algorithm for computing the treewidth and constructing a corresponding tree-decomposition for circular-arc graphs in $O(n^3)$ time.

Key words. treewidth, circular-arc graphs, tree-decomposition, complexity

AMS subject classifications. O5C85, 68Q25, 68R10

1. Introduction. The notion of the treewidth of a graph has a large number of applications in many areas, like algorithmic graph theory, VLSI design, and so forth [Ar85]. Recently there has been a growing interest in the study of the treewidth and the tree-decomposition of a graph from the algorithmic point of view.

Robertson and Seymour, in their classic series of papers on graph minors, introduced the notion of tree-decomposition and treewidth of a graph [RS86]. Most of their results are existential in nature. Several classes of problems that are NP-complete for an arbitrary graph admit polynomial-time solutions on graphs with bounded treewidth [ALS88]. Bodlaender [Bo88], [Bo89], [Bo93] has done extensive studies on the sequential and parallel algorithmic aspects of graphs with bounded treewidth.

Computing the treewidth and the corresponding tree-decomposition is known to be NP-complete for general graphs [ACP87]. For fixed k, the problem of determining whether the treewidth of a given graph is at most k can be solved in polynomial time with dynamic programming [ACP87], and in $O(n^2)$ time with graph minor theory [RS86]. The only known[1] algorithm for computing the treewidth of special classes of graphs is for cographs [BM90]. In this paper, we show that the treewidth of circular-arc graphs and the corresponding tree-decomposition can be found in $O(n^3)$ time. We do this by showing that among all possible tree-decompositions of the given circular-arc graph, only a small fraction need be considered to find one with minimum tree-width, namely, those corresponding to the planar triangulations of a certain circuit—and such tree-decompositions can be investigated quite easily with a polynomial-time algorithm.

The organization of the rest of the paper is as follows. In §2 basic definitions and preliminary results are given. Section 3 contains constructions, lemmas, and theorems that are essential to the proof of correctness of the algorithm. In §4 the actual algorithm and its analysis are presented. Section 5 contains concluding remarks.

* Received by the editors January 30, 1991; accepted for publication (in revised form) September 13, 1993. A preliminary version of this paper appeared in WADS'91. This research was conducted at the Indian Institute of Technology, Madras, as part of an ongoing project headed by C. Pandu Rangan.

† Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 (koods@theory.lcs.mit.edu).

‡ Department of Computer and Information Science, Ohio State University, 2036 Neil Avenue, Columbus, Ohio 43210 (singh-k@cis.ohio-state.edu).

§ Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India 600036 (rangan@iitm.ernet.in).

[1] More results [BKK93], [HM92] in this direction have emerged since the appearance of this paper in a conference.

2. Definitions and Preliminary Results. If S is a set, then let $|S|$ denote the cardinality of S. Let $G = (V, E)$ be a graph. We also refer to the vertex set and edge set of G by $V(G)$ and $E(G)$, respectively.

A *tree-decomposition* of G is a pair $(\{X_i|i \in I\}, T = \{I, F\})$, with $\{X_i|i \in I\}$ a family of subsets of V, and T a tree with the following properties:

(1) $\bigcup_{i\in I} X_i = V$;

(2) $\forall_{(u,v)\in E}\, \exists_{i\in I}\, (u \in X_i) \bigwedge (v \in X_i)$;

(3) $\forall_{v\in V}$, the set $\{i|(i \in I) \bigwedge (v \in X_i)\}$ forms a connected subtree of T.

The *treewidth* of a tree-decomposition $(\{X_i|i \in I\}, T = \{I, F\})$, denoted by treewidth$((\{X_i|i \in I\}, T = \{I, F\}))$, is $\max_{i\in I}(|X_i| - 1)$. The treewidth of a graph G, denoted by treewidth(G), is the minimum treewidth of a tree-decomposition of G taken over all possible tree-decompositions of G.

A graph is said to be *chordal* [Go80] if it has no chordless cycle of size greater than 3. The treewidth of a graph G can be similarly defined to be one less than the smallest clique number of a chordal graph containing G.

A *circular-arc* graph [Go80], [SP89] is the intersection graph of arcs around the unit circle. Consider a unit circle with a fixed reference, say O, on it. Any point on the circumference of the circle can be uniquely identified by its distance from O along the circle, say, in the clockwise direction. Thus, we use the same symbol for the point as well as its distance from O. Let $AF = \{A_0, A_1, \ldots, A_{n-1}\}$ be a family of arcs on a unit circle, and let $G = (V, E), |V| = n, |E| = m$, be the circular-arc graph with AF as its intersection model. The arc A_i is represented by the ordered pair $(l(A_i), r(A_i))$, where $l(A_i)$ and $r(A_i)$ denote its left and right end points, respectively. The arc A_i exists on the circle as a traversal in the clockwise direction from $l(A_i)$ to $r(A_i)$, along the circumference of the circle. Further, we assume that $l(A_i) \leq l(A_{i+1}), 0 \leq i \leq n-2$. We represent arcs in AF by uppercase letters and the corresponding vertices in G by lowercase letters. (See Figs. 1 and 2.)

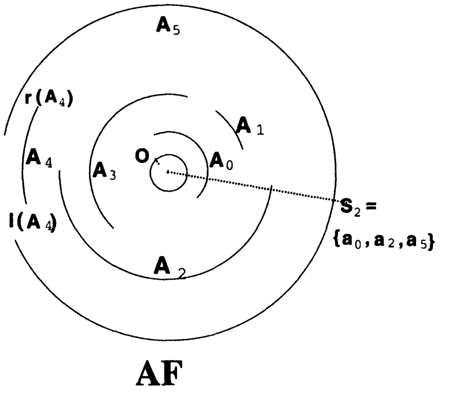


FIG. 1. *An intersection model.*

Unless otherwise stated, we assume $G = (V, E)$ to be a circular-arc graph with an arbitrary tree-decomposition $(\{X_v|v \in V(T)\}, T)$.

In general, the arc segment (x, y) exists on the circle as a traversal in the clockwise direction from x to y. (x, y) is said to *contain* position s if either $x \leq s \leq y$, or $(x \geq y) \bigwedge (s \geq x \bigvee s \leq y)$.

Every vertex $a_i \in G$ defines a *left clique* S_i, comprising the set of vertices $\{a_j|A_j$ contains $l(A_i)\}$. An *intersection clique* Q_i for the vertex a_i is the clique made

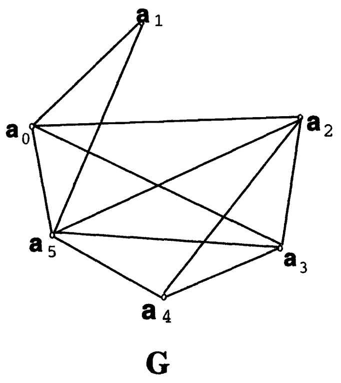


FIG. 2. *The circular-arc graph.*

up of the vertices in the set $\{a_j | a_j \in (S_i \bigcap S_{i+1})\}$. (See Fig. 3.).

$$\begin{tikzcd}
S_0 = \{a_0, a_3, a_5\} \arrow[dr] & \\
& Q_0 = \{a_0, a_5\} \\
S_1 = \{a_0, a_1, a_5\} \arrow[ur] \arrow[dr] & \\
& Q_1 = \{a_0, a_5\} \\
S_2 = \{a_0, a_2, a_5\} \arrow[ur] \arrow[dr] & \\
& Q_2 = \{a_2, a_5\} \\
S_3 = \{a_2, a_3, a_5\} \arrow[ur] \arrow[dr] & \\
& Q_3 = \{a_2, a_3\} \\
S_4 = \{a_2, a_3, a_4\} \arrow[ur] \arrow[dr] & \\
& Q_4 = \{a_3, a_4\} \\
S_5 = \{a_3, a_4, a_5\} \arrow[ur] \arrow[dr] & \\
& Q_5 = \{a_3, a_5\} \\
S_0 = \{a_0, a_3, a_5\} \arrow[ur] &
\end{tikzcd}$$

FIG. 3. *Corresponding left cliques and intersection cliques.*

Note. Unless otherwise mentioned, all indices are assumed to be modulo n.

Let

$$VQ = \bigcup_{i=0}^{n-1} Q_i.$$

Let GQ be the subgraph of G induced by VQ.

For every $v \in V$, let $ST(v)$ be the subtree of T with vertices $\{i | (i \in V(T)) \bigwedge (v \in X_i)\}$ in the tree-decomposition $(\{X_i | i \in V(T)\}, T)$. Let $ST(Y) = \bigcap_{v \in Y} ST(v)$, where Y is a subset of V.

By definition, the vertices of $ST(Y)$ also form a connected subtree of T, although it may be empty.

LEMMA 2.1 (clique containment lemma.). *In any tree-decomposition of G, for any clique K of G, $ST(K)$ is nonempty* [Bo88], [BM90].

LEMMA 2.2. *The intersection graph of a set of subtrees of a tree is chordal* [Go80].

LEMMA 2.3. *Given a cycle C of n vertices, the set of minimal chordal graphs, which contain C as a subgraph, is equal to the set of planar triangulations of C with $n - 3$ diagonal edges.*

LEMMA 2.4. *Any vertex v of G is contained in a nonempty consecutive subset $S_i, S_{i+1}, \ldots, S_j$ of the set of left cliques for some i and j, and in a consecutive subset $Q_i, Q_{i+1}, \ldots, Q_{j-1}$ of the set of intersection cliques, which is empty if $i = j$.*

LEMMA 2.5. *In any tree-decomposition of the circular-arc graph G, $ST(S_i)$ and $ST(S_{i+1})$ are subtrees of $ST(Q_i)$.*

Further, we assume no Q_i is empty, otherwise the graph reduces to an interval graph whose treewidth is trivial to compute.

COROLLARY 2.6. *$ST(Q_i) \bigcap ST(Q_{i+1})$ is nonempty for any tree-decomposition of G.*

LEMMA 2.7. *Let G be a graph. Let v be a vertex in G with degree d whose neighbours form a clique. Then* $\text{treewidth}(G) = \max(\text{treewidth}(G - v), d)$.

Proof. It is obvious that $\text{treewidth}(G) \geq \max(\text{treewidth}(G - v), d)$. To prove the upper bound, consider any chordal graph containing $G - v$; adding v yields a chordal graph containing G. (Recall that the treewidth of G is one less than the smallest clique number of a chordal graph containing G.) □

COROLLARY 2.8. $\text{Treewidth}(G) = \max(\max_{i \in I}(|S_i| - 1), \text{treewidth}(GQ))$.

Proof. From the definition of treewidth, it follows that $\text{treewidth}(G) \geq \max(\max_{i \in I} (|S_i| - 1), \text{treewidth}(GQ))$. The upper bound follows from Lemma 2.7 and the definition of GQ. □

By corollary 2.8, we see that computing treewidth(GQ) is the key to computing the treewidth of G. In what follows, we are mainly concerned with the treewidth of GQ. Note that GQ is also a circular-arc graph.

Let $G' = (V', E') = IGST(\{X_v | v \in V(T)\}, T)$ be the intersection graph of the subtrees $ST(Q_i), 0 \leq i \leq n - 1$, for a tree-decomposition $(\{X_v | v \in V(T)\}, T)$ of the circular-arc graph GQ. Let $V' = v'_0, v'_1, \ldots, v'_{n-1}$, where v'_i is the vertex corresponding to $ST(Q_i)$. From corollary 2.6, we see that $(v'_i, v'_{i+1}) \in E'$. Let CYQ be the cycle consisting of the vertices $v'_0, v'_1, \ldots, v'_{n-1}$, in that order. CYQ may also be referred to as the cycle corresponding to the intersection cliques.

LEMMA 2.9. *For any tree-decomposition $(\{X_v | v \in V(T)\}, T)$ of the circular-arc graph GQ, $IGST(\{X_v | v \in V(T)\}, T)$ (1) is chordal, and (2) contains CYQ as a subgraph.*

Proof. Part (1) follows from Lemma 2.2. (2) follows from Corollary 2.6. □

3. Constructions and Results.

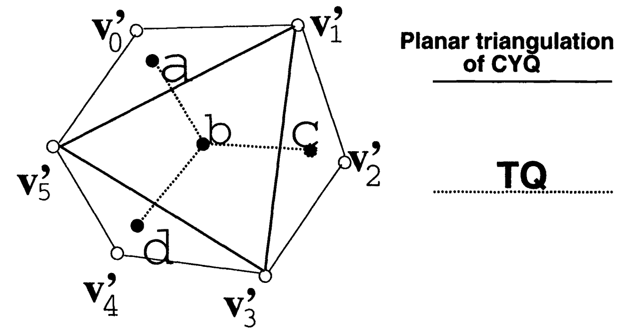
Construction 1. For a given planar triangulation of CYQ, construct a tree-decomposition $(\{X_v | v \in V(TQ)\}, TQ)$ of GQ as follows: Generate TQ to equal the dual of the planar triangulation without the vertex corresponding to the external face. Let v be any vertex of TQ and v'_i, v'_j, and v'_k be the vertices of the triangular face whose dual is the vertex v. Now, let $X_v = Q_i \bigcup Q_j \bigcup Q_k$ be the set associated with the vertex v. For each $v \in V(TQ)$ generate X_v. (See Fig. 4.)

Note. Henceforth, we assume the absence of a vertex corresponding to the external face in all references to the dual of a (planar) graph.

THEOREM 3.1. *Construction* 1 *generates a tree-decomposition of GQ.*

Proof. It is easy to see that TQ is indeed a tree.

(i) We prove that $\bigcup_{v \in TQ} X_v = VQ$. For any vertex $a_i \in VQ, a_i \in Q_i$, and since v'_i occurs in some triangular face, the corresponding dual vertex v has an associated set X_v that contains Q_i, and hence a_i. Therefore, for every vertex $a_i \in VQ$, there exists a $v \in V(TQ)$ such that $a_i \in X_v$. Further, by construction, $\forall_{v \in V(TQ)} X_v \subseteq VQ$.



$$X_a = \{a_0, a_3, a_5\}$$

$$X_b = \{a_0, a_2, a_3, a_5\}$$

$$X_c = \{a_0, a_2, a_3, a_5\}$$

$$X_d = \{a_2, a_3, a_4, a_5\}$$

FIG. 4. *Construction* 1.

Therefore,

$$(1) \qquad \bigcup_{v \in TQ} X_v = VQ.$$

(ii) We prove that $\forall_{(a_k,a_l)\in E(GQ)} \exists_{w\in V(TQ)}(a_k \in X_w) \bigwedge (a_l \in X_w)$. For any edge $(a_k, a_l) \in E(GQ), (a_k \in S_k) \bigwedge (a_l \in S_k)$, or $(a_k \in S_l) \bigwedge (a_l \in S_l)$, say $a_k, a_l \in S_k$, without loss of generality. Then $a_k \in Q_k$ and $a_l \in Q_{k-1}$. Since v'_k and v'_{k-1} are adjacent in CYQ, they occur in some triangular face, say, one whose corresponding dual vertex is w. Then w has an associated set X_w that contains Q_k and Q_{k-1}, and therefore both a_k and a_l. Therefore,

$$(2) \qquad \forall_{(a_k,a_l)\in E(GQ)} \exists_{w\in V(TQ)}(a_k \in X_w) \wedge (a_l \in X_w).$$

(iii) We prove that $\forall_{a_i\in VQ}, \{v | (v \in V(TQ)) \bigwedge (a_i \in X_v)\}$ forms a connected subtree of TQ. By Lemma 2.4, any vertex a_i of GQ occurs in a consecutive subset $Q_i, Q_{i+1}, \ldots, Q_{j-1}$ of the intersection cliques. Let a triangular face be said to intersect a set of vertices if at least one of the three vertices of the face belongs to the set. Thus we must prove that, given a planar triangulation of a cycle and a segment of the cycle, the set of vertices in the dual, whose corresponding triangular faces intersect the segment, forms a connected subtree (in the dual). The proof follows by induction on the number of vertices in the segment. If the segment contains exactly one vertex, then the corresponding subtree is simply a path in the dual tree. Assume that it holds for all segments containing k vertices. Consider a segment containing $k + 1$ vertices, numbered 1 to $k + 1$ in order along the cycle. Consider the subsegment of k vertices formed by dropping one of the end vertices, say the $k + 1$th. Corresponding to the subsegment, we have a connected subtree in the dual (by the induction hypothesis), and corresponding to the dropped end vertex, we have a path that intersects with the subtree (the intersection contains at least the dual (vertex) of the common triangular face of the kth and $k+1$th vertices in the segment). Since the dual (TQ) is a tree, we obtain a connected subtree in the dual that is the union of the subtree corresponding to the subsegment of k vertices and the path corresponding to the one end vertex of the segment. Therefore,

$$(3) \qquad \forall_{a_i \in V_Q}, \{v | (v \in V(TQ)) \wedge (a_i \in X_v)\}$$

is a connected subtree of TQ.

From (1), (2), and (3) we conclude that Construction 1 generates a tree-decomposition of GQ. □

Consider any tree-decomposition $(\{X_v | v \in V(T')\}, T')$ of GQ such that $G' = IGST(\{X_v | v \in V(T')\}, T')$, the intersection graph of the subtrees corresponding to the intersection cliques, is not a planar triangulation of CYQ. By Lemma 2.9, G' is chordal and contains CYQ. Consider any minimal subgraph SG' of G' that contains CYQ and is chordal (obviously, SG' is an edge-induced subgraph). By Lemma 2.3, SG' must be a planar triangulation of CYQ. Let $(\{X_v | v \in V(TQ')\}, TQ')$ be the tree-decomposition corresponding to SG' obtained by Construction 1.

THEOREM 3.2. *Treewidth*$(\{X_v | v \in V(TQ')\}, TQ') \leq$ *treewidth*$(\{X_v | v \in V(T')\}, T')$.

Proof. By theorem 3.1, $(\{X_v | v \in V(TQ')\}, TQ')$ is a tree-decomposition of GQ. To prove the theorem, it is sufficient to show that $\forall_{u \in V(TQ')} \exists_{v \in V(T')} : X_u \subseteq X_v$.

Consider any $u \in V(TQ')$. By Construction 1, $X_u = Q_i \bigcup Q_j \bigcup Q_k$ for some i, j, and k. (Remember that in CYQ, there is a vertex v_i' for each Q_i.) Hence, SG' contains the edges $(v_i', v_j'), (v_j', v_k')$, and (v_i', v_k'). Since SG' is only a subgraph of G', these edges are also present in G'. This implies that the sets $ST(Q_i)$, $ST(Q_j)$, and $ST(Q_k)$ pairwise overlap, but because these are subtrees, they have a nonempty intersection in T', i.e., $ST(Q_i) \bigcap ST(Q_j) \bigcap ST(Q_k)$ in T' is nonempty. Consider some vertex $v \in V(T')$ in this nonempty intersection. X_v contains Q_i, Q_j and Q_k. Therefore, $X_v \supseteq X_u$.

Hence, it has been proved that treewidth$(\{X_v | v \in V(TQ')\}, TQ') \leq$ treewidth$(\{X_v | v \in V(T')\}, T')$. □

COROLLARY 3.3. *The treewidth of GQ is equal to the minimum treewidth over all tree-decompositions $(\{X_v | v \in V(TQ)\}, TQ)$ of GQ that satisfy the following: $G' = IGST(\{X_v | v \in V(TQ)\}, TQ)$ the corresponding intersection graph of the intersection cliques, contains CYQ and is a planar triangulation of CYQ.*

Proof. It follows from Theorems 3.1 and 3.2, and the fact that $IGST(\{X_v | v \in V(T)\}, T)$ is chordal and contains CYQ for any tree-decomposition $(\{X_v | v \in V(T)\}, T)$ of GQ. □

Construction 2. Given any tree-decomposition $(\{X_v | v \in V(TQ)\}, TQ)$ of GQ such that $G' = IGST(\{X_v | v \in V(TQ)\}, TQ)$ contains CYQ and is chordal, construct a tree-decomposition of G as follows: For each $a_i \in V$, add a new vertex b_i to TQ; attach b_i by an edge to any one vertex $v \in V(TQ)$ such that X_v contains Q_{i-1} and Q_i; and set $X_{b_i} = S_i$. Let $(\{X_v | v \in V(TQ_G)\}, TQ_G)$ denote the resulting tree-decomposition. (See Fig. 5.).

It is easy to see that Construction 2 is well defined and that it generates a tree-decomposition of G. Further, by Corollary 2.8, if the tree decomposition $(\{X_v | v \in V(TQ)\}, TQ)$ achieves the minimum tree-width for GQ, then the treewidth of the tree-decomposition $(\{X_v | v \in V(TQ_G)\}, TQ_G)$ equals the treewidth of G.

4. Algorithm and its analysis. The basic algorithm now boils down to computing treewidth(GQ), and a corresponding tree-decomposition. The problem of com-

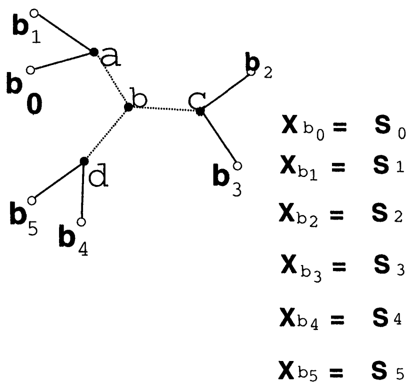


FIG. 5. *Construction* 2.

puting treewidth(GQ) can be rephrased as follows.

Given an n-gon $v_0, v_1, \ldots, v_{n-1}$, and sets A_i associated with each vertex v_i such that every element occurs in a consecutive subset $A_i, A_{i+1}, \ldots, A_j$, define the cost of a triangle with vertices v_i, v_j, and v_k to be $|A_i \bigcup A_j \bigcup A_k| - 1$. Define the cost of a planar triangulation to be the maximum cost over all triangular faces in the triangulation. The problem (restated) involves finding the minimum cost over all planar triangulations, and the corresponding planar triangulation (refer to Corollary 3.3).

Specifically, in this case it is required to find the minimum-cost planar triangulation of the n-gon $v'_0, v'_1, \ldots, v'_{n-1}$, with the set associated with v'_i being Q_i. This is done by a dynamic programming approach. $TWCYQ_j(i)$ is defined to be the minimum cost over all planar triangulations of the subpolygon $v'_i, v'_{i+1}, \ldots, v'_{i+j-1}$. Therefore, treewidth($GQ$) $= TWCYQ_n(0)$.

ALGORITHM.

Input $(l(A_i), r(A_i))$ representation of the vertices a_i of the circular-arc graph G.

Steps

1. Construct $S_i, 0 \leq i \leq n-1$ (in sorted order).
2. Construct $Q_i, 0 \leq i \leq n-1$ (in sorted order).
3. Compute treewidth(GQ) $= TWCYQ_n(0)$ using dynamic programming, and find a corresponding planar triangulation.
4. Employ Construction 1 on the planar triangulation to obtain the tree-decomposition of GQ.
5. Employ Construction 2 (using the tree-decomposition of GQ obtained from step 4) to obtain the tree-decomposition of G and compute the corresponding treewidth.

Output Treewidth(G) and corresponding tree-decomposition.

Details of step 3.

3.1. For all $i, 0 < i < n-1$, compute

$$TWCYQ_1(i) = TWCYQ_2(i) = 0,$$

$$TWCYQ_3(i) = |Q_i \cup Q_{i+1} \cup Q_{i+2}| - 1.$$

3.2. For $j = 4$ to n, and
for $i = 0$ to $n - 1$, compute

$$TWCYQ_j(i) = \min_{k=1}^{j-2}\{\max(TWCYQ_{k+1}(i), TWCYQ_{j-k}(i+k), |Q_i \cup Q_{i+k} \cup Q_{i+j-1}| - 1)\}.$$

To actually find the corresponding planar triangulation, pointers and additional book-keeping, information could be stored so that, by retracing, the optimum triangulation can be constructed.

Proof of correctness. It follows from Theorems 3.1 and 3.2 and Corollary 3.3. □

Time complexity. The dynamic programming employed in step 3 itself takes $O(n^3)$ time. But the bottleneck is finding all the $|Q_i \bigcup Q_{i+k} \bigcup Q_{i+j-1}|$. Since the union in the innermost loop takes $O(n)$ time, naively it would seem that step 3 takes $O(n^4)$ time. However, finding all the $|Q_i \bigcup Q_{i+k} \bigcup Q_{i+j-1}|$ can be done faster than $O(n^4)$. Note that the sets Q_i and the elements they contain can be represented as a circular list of events. The events represent the start of an element, the end of an element, and a set. An element is considered to belong to all those sets whose event lies in the clockwise traversal from the start event, to the end event of this element. To show that all the $|Q_i \bigcup Q_{i+k} \bigcup Q_{i+j-1}|$ can be computed in $O(n^3)$ time, all we must show is that for any fixed set B, all the $|B \bigcup Q_i|$ can be computed in $O(n)$ time. This is done as follows: Remove all events corresponding to elements in B from the list, then do a sweep around the circle maintaining a counter that represents the arcs of $Q_i - B$ covering the point you are at, decrement the counter when you hit the end of an arc, and increment it at the start of an arc. When you arrive at the event corresponding to a set, output the counter plus $|B|$.

5. Final remarks. In this paper, we presented an algorithm to compute the treewidth of a circular-arc graph in $O(n^3)$ time. Since our algorithm simply uses straightforward dynamic programming, we conjecture that it is possible to improve the time complexity substantially. We note that there is an $O(n \log n)$ algorithm for the following problem (which is syntactically very similar to the one we solve).

Given an n-gon $v_0, v_1, \ldots, v_{n-1}$, and numbers A_i associated with each vertex v_i, define the cost of a triangle with vertices v_i, v_j and v_k to be $A_i \times A_j \times A_k$. Define the cost of a planar triangulation to be the sum over all triangular faces in the triangulation. The problem involves finding the minimum cost over all planar triangulations and the corresponding planar triangulation.

The fastest known algorithm for this problem was the straightforward $O(n^3)$ dynamic programming algorithm, until Hu and Shing devised an extremely clever $O(n \log n)$ algorithm [HS80], [Ya80]. So it is possible that a closer analysis of the underlying problem will yield a faster algorithm. Another interesting direction is to develop an algorithm to compute the treewidth for a substantially larger class of graphs.

Acknowledgments. The authors thank the anonymous referees for suggesting the many improvements and corrections that have gone a long way toward improving the presentation of this paper.

REFERENCES

[Ar85] S. Arnborg, *Efficient algorithms for combinatorial problems on graphs with bounded decomposability—A survey*, BIT, 25 (1985), pp. 2–23.

[ACP87] S. Arnborg, D. G. Corneil, and A. Proskurowski, *Complexity of finding embeddings in a k-tree*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 277–284.

[ALS88] S. Arnborg, J. Lagergren, and D. Seese, *Problems easy for tree-decomposable graphs*, in Proc. 15th ICALP, LNCS 317 (1988), pp. 38–51.

[Bo88] H. L. Bodlaender, *NC-algorithms for graphs with small treewidth*, in Proc. WG'88, LNCS 344 (1988), pp. 1–10.

[Bo89] ———, *On linear time minor tests and depth first search*, In Proc. WADS'89, LNCS 382 (1989), pp. 577–590.

[Bo93] ———, *A linear time algorithm for finding tree-decompositions of small treewidth*, in Proc. 25th Annual ACM Symposium on the Theory of Computing, 1993, pp. 226–234.

[BM90] H. L. Bodlaender and R. H. Mohring, *The pathwidth and treewidth of cographs*, in Proc. SWAT'90, LNCS 447 (1990), pp. 301–309.

[BKK93] H. L. Bodlaender, T. Kloks and D. Kratsch, *Treewidth and pathwidth of permutation graphs*, in Proc. 20th ICALP, 1993, to appear.

[Go80] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[HM92] M. Habib and R. H. Mohring, *Treewidth of cocomparability graphs and a new order-theoretic parameter*, Tech. Report 336/1992, Technische Univ. Berlin, 1992.

[HS80] T. C. Hu and M. T. Shing, *Some theorems about matrix multiplication* in Proc. 21st Annual IEEE Symposium on Foundations of Computer Science, 1980, pp. 28–35.

[RS86] N. Robertson and P. Seymour, *Graph minors.* II. *Algorithmic aspects of treewidth*, J. Algorithms, 7 (1986), pp. 309–322.

[SP89] A. Srinivasa Rao and C. Pandu Rangan, *Linear algorithms for parity path and two path problems on circular-arc graphs*, in Proc. WADS'89, LNCS 382 (1989), pp. 267–290.

[Ya80] F. F. Yao, *Efficient dynamic programming using quadrangle inequalities*, in Proc. 12th Annual ACM Symposium on the Theory of Computing, 1980, pp. 429–435.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 656-666, November 1994

NEW $\frac{3}{4}$-APPROXIMATION ALGORITHMS FOR THE MAXIMUM SATISFIABILITY PROBLEM*

MICHEL X. GOEMANS† AND DAVID P. WILLIAMSON‡

Abstract. Yannakakis recently presented the first $\frac{3}{4}$-approximation algorithm for the Maximum Satisfiability Problem (MAX SAT). His algorithm makes nontrivial use of solutions to maximum flow problems. New, simple $\frac{3}{4}$-approximation algorithms that apply the probabilistic method/randomized rounding to the solution to a linear programming relaxation of MAX SAT are presented. It is shown that although standard randomized rounding does not give a good approximate result, the best solution of the two given by randomized rounding and a well-known algorithm of Johnson is always within $\frac{3}{4}$ of the optimal solution. It is further shown that an unusual twist on randomized rounding also yields $\frac{3}{4}$-approximation algorithms. As a by-product of the analysis, a tight worst-case analysis of the relative duality gap of the linear programming relaxation is obtained.

Key words. approximation algorithm, maximum satisfiability, randomized rounding, probabilistic method, performance guarantee, linear programming relaxations

AMS subject classifications. 90C27, 68Q25, 90C05

1. Introduction. An instance of the Maximum Satisfiability Problem (MAX SAT) is defined by a collection $\mathcal{C}$ of boolean clauses, where each clause is a disjunction of literals drawn from a set of variables $\{x_1, x_2, \ldots, x_n\}$. A *literal* is either a variable x or its negation $\bar{x}$. In addition, for each clause $C_j \in \mathcal{C}$, there is an associated nonnegative weight w_j. An optimal solution to a MAX SAT instance is an assignment of truth values to variables $x_1, \ldots, x_n$ that maximizes the sum of the weight of the satisfied clauses (i.e., clauses with at least one true literal). MAX SAT is known to be NP-complete, even when each clause contains at most two literals (sometimes called MAX 2SAT) [4]. Hence there is unlikely to be any polynomial-time algorithm that can solve MAX SAT optimally.

Many people, however, have proposed *α-approximation algorithms* for MAX SAT. An α-approximation algorithm for MAX SAT is a polynomial-time algorithm which, for every instance, produces a truth assignment with weight at least α times the weight of an optimal solution. Johnson [7] demonstrates a $\frac{1}{2}$-approximation algorithm, which is also an $(1 - \frac{1}{2^k})$-approximation algorithm when each clause contains at least k literals. In particular, if $k \geq 2$ the performance guarantee is at least $\frac{3}{4}$. Lieberherr and Specker [9] give a $\frac{\sqrt{5}-1}{2}$-approximation algorithm ($\frac{\sqrt{5}-1}{2} = 0.618\ldots$) when the clause set does not contain both clauses x_i and $\bar{x}_i$ for any i. Kohli and Krishnamurti [8] present a randomized algorithm whose solution has expected weight at least $\frac{2}{3}$ of optimal. Yannakakis recently improved on these results by showing a $\frac{3}{4}$-approximation algorithm [13]. Yannakakis' algorithm transforms a MAX SAT instance into an equivalent instance (in terms of approximability) which does not contain any unit clauses (i.e., clauses with only one literal). In conjunction with Johnson's algorithm, this

* Received by the editors October 19, 1992; accepted for publication (in revised form) October 8, 1993.

† Department of Mathematics, Massachusetts Institute of Technology, Room 2-372, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139 (goemans@math.mit.edu). This author's research was supported in part by Air Force Office of Scientific Research contract AFOSR-89-0271 and DARPA contract N00014-89-J-1988.

‡ School of Operations Research and Industrial Engineering, Cornell University, 237 ETC Building, Ithaca, New York 14853 (dpw@cs.cornell.edu). This author's research was supported by a National Science Foundation Graduate Fellowship and DARPA contract N00014-89-J-1988.

leads to the improved performance guarantee. The algorithm uses maximum flow computations in an elegant way to transform MAX 2SAT instances. However, the transformation becomes more complicated when general clauses are introduced.

The purpose of this article is to present new $\frac{3}{4}$-approximation algorithms which are conceptually simple for all MAX SAT instances. The algorithms presented here apply the technique of randomized rounding (Raghavan and Thompson [11], [10]) to the solution of a single linear program that is a linear programming relaxation of a formulation for the MAX SAT problem. However, a straightforward application of the technique does not yield a $\frac{3}{4}$-approximation algorithm. We surmount this difficulty in two ways: by combining randomized rounding with Johnson's algorithm and by using an interesting variation of the standard randomized rounding technique.

The article is structured as follows. In §2, Johnson's algorithm is reviewed in terms of the probabilistic method. In §3, we show that a straightforward application of randomized rounding to a linear programming relaxation of MAX SAT leads to a $(1 - \frac{1}{e})$-approximation algorithm ($1 - 1/e = 0.632\ldots$). The algorithm that selects the better of the two solutions given by randomized rounding and Johnson's algorithm is shown to be a $\frac{3}{4}$-approximation algorithm in §4. In §5, we describe a class of $\frac{3}{4}$-approximation algorithms for MAX SAT based on a variant of randomized rounding. We conclude with a few remarks in §6.

2. Johnson's algorithm and the probabilistic method. Suppose we independently and randomly set each variable x_i to be true with probability p_i. Then the expected weight of clauses satisfied by this probabilistic assignment is

$$\hat{W} = \sum_{C_j \in C} w_j \left(1 - \prod_{i \in I_j^+} (1 - p_i) \prod_{i \in I_j^-} p_i \right),$$

where I_j^+ (resp., I_j^-) denotes the set of variables appearing unnegated (resp., negated) in C_j. The probabilistic method specifies that there must exist an assignment of truth values to the variables whose weight is at least this expected value. In fact, the method of conditional probabilities (see Alon and Spencer [1], p. 223) can be applied to find such an assignment deterministically in polynomial time. In the method of conditional probabilities, the value for the ith variable is determined in the ith iteration: given the values of $x_1, \ldots, x_{i-1}$, calculate the expected weight of clauses satisfied by the probabilistic assignment, given the current assignment to $x_1, \ldots, x_{i-1}$ and the assignment $x_i = 1$. Then calculate the expected weight given the assignment to $x_1, \ldots, x_{i-1}$ and $x_i = 0$. The variable x_i is assigned the value that maximizes the conditional expectation. Since each conditional expectation can be calculated in polynomial time, the overall algorithm takes polynomial time, and as asserted above, the assignment produced has weight at least $\hat{W}$.

As interpreted by Yannakakis [13], Johnson's algorithm essentially sets $p_i = \frac{1}{2}$ for all i and uses the method of conditional probabilities. It is not hard to see that for this choice of p_i,

$$\hat{W} \geq \sum_{C_j \in C} \left(1 - \frac{1}{2}\right) w_j = \frac{1}{2} \sum_{C_j \in C} w_j.$$

Since the optimum assignment can have weight at most $\sum_j w_j$, this proves that Johnson's algorithm is a $\frac{1}{2}$-approximation algorithm. Moreover, if all clauses have at least

k literals then

$$\hat{W} \geq \left(1 - \frac{1}{2^k}\right) \sum_{C_j \in \mathcal{C}} w_j,$$

implying that Johnson's algorithm is a $(1 - \frac{1}{2^k})$-approximation algorithm for this restricted class of instances.

3. A $(1 - \frac{1}{e})$-approximation algorithm.

Consider the following integer program:

$$\begin{array}{llll} & \text{Max} & \displaystyle\sum_{C_j \in \mathcal{C}} w_j z_j & \\ & \text{subject to:} & & \\ (IP) & & \displaystyle\sum_{i \in I_j^+} y_i + \sum_{i \in I_j^-} (1 - y_i) \geq z_j & \forall C_j \in \mathcal{C} \\ & & y_i \in \{0, 1\} & 1 \leq i \leq n \\ & & 0 \leq z_j \leq 1 & \forall C_j \in \mathcal{C}. \end{array}$$

By associating $y_i = 1$ with x_i set true, $y_i = 0$ with x_i set false, $z_j = 1$ with clause C_j satisfied, and $z_j = 0$ with clause C_j not satisfied, the integer program (IP) exactly corresponds to the MAX SAT problem, and its optimal value Z_{IP}^* is equal to the optimal value of the MAX SAT problem. We can now consider the linear programming relaxation of (IP) formed by replacing the $y_i \in \{0, 1\}$ constraints with the constraints $0 \leq y_i \leq 1$. Call this linear program (LP). Obviously the optimal value of (LP) is an upper bound on the optimal value of (IP); that is, $Z_{LP}^* \geq Z_{IP}^*$. Whenever there are no unit clauses, the solution $y_i = \frac{1}{2}$ for all i and $z_j = 1$ for all j, which is of value $\sum_{C_j \in \mathcal{C}} w_j$, is optimal, independent of the weights w_j. Hence, the relaxation is vacuous in this case. However, when there are unit clauses (the bad case for Johnson's algorithm), we shall show in this and later sections that this relaxation provides some useful information.

We now show that by using randomized rounding in a straightforward fashion we obtain a $(1 - \frac{1}{e})$-approximation algorithm for MAX SAT. This algorithm consists of two simple steps. The first step is to solve the linear program (LP). Let (y^*, z^*) be an optimal solution. The second step is to apply the method of conditional probabilities with $p_i = y_i^*$ for all i to derive an assignment. By using Tardos' algorithm [12] to solve (LP), the algorithm runs in strongly polynomial time since the constraint matrix of (LP) has all entries in $\{-1, 0, 1\}$.

The proof of the performance guarantee of $1 - \frac{1}{e}$ is similar to the approach described in §2 although the expected weight $\hat{W}$ of a random truth assignment is not compared to $\sum_{C_j \in \mathcal{C}} w_j$ but rather to Z_{LP}^*. Notice that if

$$1 - \prod_{i \in I_j^+} (1 - y_i) \prod_{i \in I_j^-} y_i \geq \alpha z_j$$

for any feasible solution (y, z) to (LP) and for any clause C_j then

$$\hat{W} = \sum_{C_j} w_j \left\{ 1 - \prod_{i \in I_j^+} (1 - p_i) \prod_{i \in I_j^-} p_i \right\}$$

$$= \sum_{C_j} w_j \left\{ 1 - \prod_{i \in I_j^+} (1 - y_i^*) \prod_{i \in I_j^-} y_i^* \right\}$$
$$\geq \alpha \sum_{C_j} w_j z_j^* = \alpha Z_{LP}^* \geq \alpha Z_{IP}^*,$$

implying that the resulting algorithm is an α-approximation algorithm.

LEMMA 3.1. *For any feasible solution (y, z) to (LP) and for any clause C_j with k literals, we have*

$$1 - \prod_{i \in I_j^+} (1 - y_i) \prod_{i \in I_j^-} y_i \geq \beta_k z_j$$

where

$$\beta_k = 1 - \left(1 - \frac{1}{k}\right)^k .$$

This and subsequent proofs use the following simple results. To show that a concave function $f(x)$ satisfies $f(x) \geq ax + b$ over the interval $[l, u]$, one only needs to show it for the endpoints of the interval, namely $f(l) \geq al + b$ and $f(u) \geq au + b$. We shall also rely on the arithmetic/geometric mean inequality which states that

$$\frac{a_1 + a_2 + \cdots + a_k}{k} \geq \sqrt[k]{a_1 a_2 \cdots a_k},$$

for any collection of nonnegative numbers $a_1, a_2, \ldots, a_k$.

Proof. We can assume without loss of generality that all variables in the clause are unnegated. Indeed, if x_i appears negated in clause C_j, one can replace x_i by its negation $\overline{x}_i$ in every clause and also replace y_i by $1 - y_i$ without affecting the feasibility of (LP) or the claim stated in the lemma. We thus assume that the clause is $x_1 \vee \cdots \vee x_k$ with associated constraint $y_1 + \cdots + y_k \geq z_j$. We need to prove that

$$1 - \prod_{i=1}^{k} (1 - y_i) \geq \beta_k z_j .$$

Applying the arithmetic/geometric mean inequality to $\{1 - y_i\}$ and using the constraint on z_j, we obtain that

$$1 - \prod_{i=1}^{k} (1 - y_i) \geq 1 - \left(1 - \frac{\sum_{i=1}^{k} y_i}{k}\right)^k$$
$$\geq 1 - \left(1 - \frac{z_j}{k}\right)^k .$$

Since $f(z_j) = 1 - \left(1 - \frac{z_j}{k}\right)^k$ is a concave function and since $f(0) = 0$ and $f(1) = \beta_k$, we derive that

$$1 - \left(1 - \frac{z_j}{k}\right)^k \geq \beta_k z_j ,$$

proving the desired result. □

Since β_k is decreasing with k, Lemma 3.1 and the discussion that precedes it show that this simple algorithm is a β_k-approximation algorithm for the class of MAX SAT instances with at most k literals per clause. In particular, it is a $\frac{3}{4}$-approximation algorithm for MAX 2SAT and a $\left(1-\frac{1}{e}\right)$-approximation algorithm for MAX SAT in general, since $\lim_{k\to\infty}\left(1-\frac{1}{k}\right)^k=\frac{1}{e}$.

In a certain sense, the analysis we have just performed cannot be improved. Consider the MAX SAT instance consisting of the clauses $C_j : \bigvee_{i\neq j} x_i$ for $j=1,\ldots,n$ with weight n and the clauses $C_{n+j} : \bar{x}_j$ for $j=1,\ldots,n$ with weight 1. One can show that the *unique* optimum solution to (LP) is given by $y_i^* = \frac{1}{n-1}$ for all $i=1,\ldots,n$ and

$$z_j^* = \begin{cases} 1 & j \leq n \\ 1-\frac{1}{n-1} & j > n. \end{cases}$$

One can further show that

$$\lim_{n\to\infty}\frac{\hat{W}}{Z_{IP}^*} = \lim_{n\to\infty}\frac{\hat{W}}{Z_{LP}^*} = 1-\frac{1}{e},$$

and thus the inequality $\hat{W} \geq \left(1-\frac{1}{e}\right)Z_{IP}^*$ is tight. However, applying the method of conditional probabilities to this optimum (LP) solution yields the optimum truth assignment.

4. A simple $\frac{3}{4}$-approximation algorithm.

In §2, we have shown that Johnson's algorithm is a $\frac{3}{4}$-approximation algorithm when all clauses contain *at least* 2 literals, while in the previous section, we have presented a $\frac{3}{4}$-approximation algorithm when all clauses contain *at most* 2 literals (i.e., for MAX 2SAT instances). In this section, we show that a $\frac{3}{4}$-approximation algorithm can be obtained by choosing the best truth assignment between the two output by Johnson's algorithm and the algorithm of the previous section. More formally, we have the following result.

THEOREM 4.1. *Let $\hat{W}_1$ denote the expected weight corresponding to $p_i=\frac{1}{2}$ for all i and let $\hat{W}_2$ denote the expected weight corresponding to $p_i = y_i^*$ for all i where (y^*, z^*) is an optimum solution to the (LP) relaxation. Then*

$$\max(\hat{W}_1,\hat{W}_2) \geq \frac{\hat{W}_1+\hat{W}_2}{2} \geq \frac{3}{4}Z_{LP}^*.$$

Proof. The first inequality is trivially satisfied. Let $\mathcal{C}^k$ denote the set of clauses with exactly k literals. From §2, we know that

$$\hat{W}_1 = \sum_{k\geq 1}\sum_{C_j\in\mathcal{C}^k}\alpha_k w_j \geq \sum_{k\geq 1}\sum_{C_j\in\mathcal{C}^k}\alpha_k w_j z_j^*,$$

where $\alpha_k = \left(1-\frac{1}{2^k}\right)$. On the other hand, Lemma 3.1 implies that

$$\hat{W}_2 \geq \sum_{k\geq 1}\sum_{C_j\in\mathcal{C}^k}\beta_k w_j z_j^*,$$

where

$$\beta_k = 1-\left(1-\frac{1}{k}\right)^k.$$

As a result,

$$\frac{\hat{W}_1+\hat{W}_2}{2} \geq \sum_{k\geq 1}\sum_{C_j\in\mathcal{C}^k} \frac{\alpha_k+\beta_k}{2} w_j z_j^*.$$

Clearly, $\alpha_1+\beta_1=\alpha_2+\beta_2=\frac{3}{2}$ while, for $k\geq 3$, $\alpha_k+\beta_k\geq\frac{7}{8}+1-\frac{1}{e}\geq\frac{3}{2}$. Therefore,

$$\frac{\hat{W}_1+\hat{W}_2}{2} \geq \sum_{k\geq 1}\sum_{C_j\in\mathcal{C}^k} \frac{3}{4} w_j z_j^* = \frac{3}{4} Z_{LP}^*. \qquad \square$$

The previous theorem also demonstrates that the following algorithm is a $\frac{3}{4}$-approximation algorithm: with probability $\frac{1}{2}$, set the vector p of probabilities to be either $p_i=\frac{1}{2}$ for all i or $p_i=y_i^*$ for all i, and apply the method of conditional probabilities. In this scheme, x_i is set true with probability $\frac{1}{4}+\frac{1}{2}y_i^*$ but this algorithm does not fit in the framework described in §2 since the x_i's are not set *independently.*

5. A class of $\frac{3}{4}$-approximation algorithms. The standard randomized rounding scheme of §3 can be modified to lead directly to $\frac{3}{4}$-approximation algorithms. For this purpose, instead of using $p_i=y_i^*$ for all i, we let $p_i=f(y_i^*)$ for some carefully selected function $f:[0,1]\rightarrow[0,1]$ and, as before, apply the method of conditional probabilities. Possible choices of f are discussed below. As far as we know, this is the first application of randomized rounding in which the probabilities p_i are not identical to or scaled versions of the linear program solutions y_i^*.

As in §3, if we can show that

$$(1) \qquad 1-\prod_{i\in I_j^+}(1-f(y_i))\prod_{i\in I_j^-}f(y_i)\geq\frac{3}{4}z_j$$

for any feasible solution (y,z) to (LP) and for any clause C_j, then the resulting algorithm is a $\frac{3}{4}$-approximation algorithm. Inequality (1) together with the constraints on z_j motivates the following definition.

DEFINITION 5.1. *A function $f:[0,1]\rightarrow[0,1]$ has property $\frac{3}{4}$ if*

$$1-\prod_{i=1}^{l}(1-f(y_i))\prod_{i=l+1}^{k}f(y_i)\geq\frac{3}{4}\min\left(1,\sum_{i=1}^{l}y_i+\sum_{i=l+1}^{k}(1-y_i)\right)$$

for any k,l with $k\geq l$ and any $y_1,\ldots,y_k\in[0,1]$.

By the discussion of §3, any function f with property $\frac{3}{4}$ induces a $\frac{3}{4}$-approximation algorithm. The following theorems show that not only do there exist functions with property $\frac{3}{4}$ but also that there is some flexibility in choosing such a function.

THEOREM 5.2. *Any function f satisfying*

$$1-4^{-y}\leq f(y)\leq 4^{y-1}$$

for all $y\in[0,1]$ has property $\frac{3}{4}$.

THEOREM 5.3. *The linear function $f_\alpha(y)=\alpha+(1-2\alpha)y$, where*

$$2-\frac{3}{\sqrt[3]{4}}\leq\alpha\leq\frac{1}{4},$$

has property $\frac{3}{4}(2 - \frac{3}{\sqrt[3]{4}} \approx .11)$.

THEOREM 5.4. *The function*

$$f(y) = \begin{cases} \frac{3}{4}y + \frac{1}{4} & \textit{if}\, 0 \le y \le \frac{1}{3} \\ \frac{1}{2} & \textit{if}\, \frac{1}{3} \le y \le \frac{2}{3} \\ \frac{3}{4}y & \textit{if}\, \frac{2}{3} \le y \le 1 \end{cases}$$

has property $\frac{3}{4}$.

The possible choices for f following from Theorems 5.2–5.4 are depicted in Fig. 1. Before proving these theorems, we would like to make a few remarks regarding the functions with property $\frac{3}{4}$ given in these theorems.

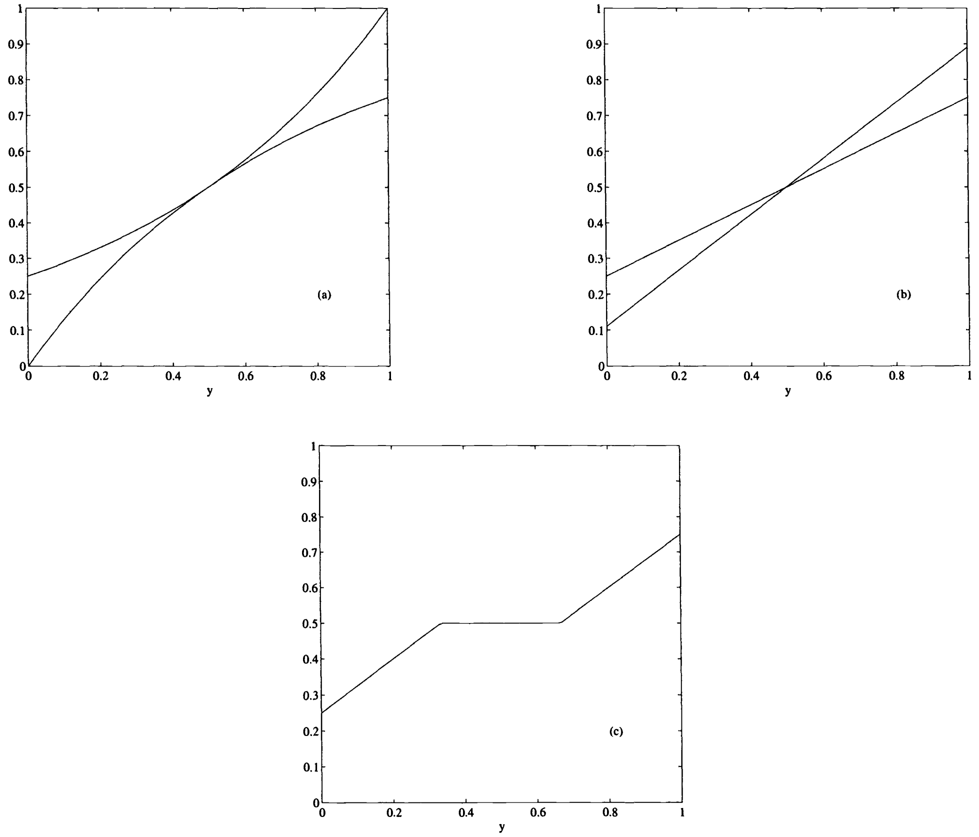


FIG. 1. *Functions with property* $\frac{3}{4}$ *from Theorems* (a) 5.2, (b) 5.3, *and* (c) 5.4.

1. There exist functions satisfying the conditions of Theorem 5.2 since, by the arithmetic/geometric mean inequality, $\frac{(4^{-y}+4^{y-1})}{2} \ge \frac{1}{2}$, i.e., $1 - 4^{-y} \le 4^{y-1}$.

2. By Theorem 5.2, there exist functions f with property $\frac{3}{4}$ for which $f(1) = 1$ and $f(0) = 0$. This is the case, for example, for

$$f(y) = \begin{cases} 4^{y-1} & \text{if } y \ge \frac{1}{2} \\ 1 - 4^{-y} & \text{if } y < \frac{1}{2}. \end{cases}$$

The property $f(1) = 1$ and $f(0) = 0$ implies that the randomized rounding sets the value of a variable x_i deterministically according to y_i^* when $y_i^* \in \{0, 1\}$.

3. For $l = k$ and $y_1 = \cdots = y_k = \frac{1}{k}$, notice that any function f with property $\frac{3}{4}$ must satisfy

$$f\left(\frac{1}{k}\right) \geq 1 - 4^{-1/k}$$

for all integers $k \geq 1$. This partially explains the choice of the lower bound in Theorem 5.2. The upper bound can be similarly obtained by considering $l = 0$.

4. Although $f(y) = y$ is not a function with property $\frac{3}{4}$, there exist linear functions with property $\frac{3}{4}$ as demonstrated in Theorem 5.3.

5. The linear function $f_{1/4}(y)$ of Theorem 5.3 corresponds to setting x_i to be true with probability $\frac{1}{4} + \frac{1}{2}y_i^*$. However, the resulting algorithm differs from the one mentioned after Theorem 4.1 since in the latter the x_i's are not set independently of each other.

6. The function described in Theorem 5.4 is the "closest" to Johnson's scheme in the sense that, for any $y \in [0, 1]$, the function f of Theorem 5.4 minimizes $|f(y) - \frac{1}{2}|$ over all functions with property $\frac{3}{4}$. Indeed, by considering the case $k = 1$, $l = 0$ or 1, one derives that any function f with property $\frac{3}{4}$ satisfies

$$\frac{3}{4}y \leq f(y) \leq \frac{3}{4}y + \frac{1}{4}$$

for any $y \in [0, 1]$.

Our proofs of Theorems 5.2–5.4 use similar ideas, but the proof of Theorem 5.4 is more tedious than the others since we have to differentiate between several cases. For this reason, we have omitted its proof, but the reader can find it in an earlier version of this paper [6]. To prove the theorems, we use the following lemma to restrict our attention to the case $l = k$ (corresponding to a clause with no negated variable).

LEMMA 5.5. *Let $g : [0, 1] \to [0, 1]$ be a function satisfying*

$$1 - \prod_{i=1}^{k}(1 - g(y_i)) \geq \frac{3}{4} \min\left(1, \sum_{i}^{k} y_i\right) \tag{2}$$

for all k and all $y_1, y_2, \ldots, y_k \in [0, 1]$. Consider any function $f : [0, 1] \to [0, 1]$ satisfying

$$g(y) \leq f(y) \leq 1 - g(1 - y) \tag{3}$$

for all $y \in [0, 1]$. Then f has property $\frac{3}{4}$.

Proof. Consider any k, l with $k \geq l$ and any $y_1, \ldots, y_k \in [0, 1]$. Then

$$\begin{aligned} 1 - \prod_{i=1}^{l}(1 - f(y_i)) \prod_{i=l+1}^{k} f(y_i) &\overset{(3)}{\geq} 1 - \prod_{i=1}^{l}(1 - g(y_i)) \prod_{i=l+1}^{k} (1 - g(1 - y_i)) \\ &= 1 - \prod_{i=1}^{k}(1 - g(y_i')) \\ &\overset{(2)}{\geq} \frac{3}{4} \min\left(1, \sum_{i=1}^{k} y_i'\right) \\ &= \frac{3}{4} \min\left(1, \sum_{i=1}^{l} y_i + \sum_{i=l+1}^{k} (1 - y_i)\right), \end{aligned}$$

where $y'_i = y_i$ for $i = 1, \ldots, l$ and $y'_i = 1 - y_i$ for $i = l+1, \ldots, k$. □

Proof of Theorem 5.2. By Lemma 5.5, we only need to show that $g(y) = 1 - 4^{-y}$ satisfies (2). We have

$$\begin{aligned} 1 - \prod_{i=1}^{k}(1 - g(y_i)) &= 1 - \prod_{i=1}^{k} 4^{-y_i} \\ &= 1 - 4^{-\sum_{i=1}^{k} y_i} \\ &= 1 - 4^{-Y} = g(Y) \end{aligned}$$

where $Y = \sum_{i=1}^{k} y_i$. Since $g(Y)$ is increasing with Y, in order to prove (2) we only need to show that $g(Y) \geq \frac{3}{4}Y$ for any $Y \in [0,1]$. This follows from the concavity of $g(Y)$ and the facts that $g(0) = 0$ and $g(1) = \frac{3}{4}$. □

Proof of Theorem 5.3. Suppose $2 - \frac{3}{\sqrt[3]{4}} \leq \alpha \leq \frac{1}{4}$. Since $f_\alpha(y) = \alpha + (1 - 2\alpha)y$ satisfies $1 - f_\alpha(1-y) = f_\alpha(y)$, we only need to show that $f_\alpha(y)$ satisfies (2) in order to use Lemma 5.5. We have

$$\begin{aligned} 1 - \prod_{i=1}^{k}(1 - f_\alpha(y_i)) &= 1 - \prod_{i=1}^{k}(1 - \alpha - (1-2\alpha)y_i) \\ &\geq 1 - \left(1 - \alpha - (1-2\alpha)\frac{\sum_{i=1}^{k} y_i}{k}\right)^k \end{aligned}$$

by the arithmetic/geometric mean inequality. Letting $y = (\sum_{i=1}^{k} y_i)/k$, we need to show that

$$1 - (1 - \alpha - (1-2\alpha)y)^k \geq \frac{3}{4}\min(1, ky), \tag{4}$$

for any $y \in [0,1]$. Since the left-hand side is increasing with y, we can restrict our attention to $y \in [0, \frac{1}{k}]$. Furthermore, since it is concave in y, we can just check (4) for $y = 0$ (for which it is trivially satisfied) and for $y = \frac{1}{k}$. For this latter value, we need to prove that

$$\frac{1}{4} \geq \left(1 - \alpha - (1-2\alpha)\frac{1}{k}\right)^k, \tag{5}$$

for any integer $k \geq 1$. For $k = 1$, (5) reduces to $\alpha \leq \frac{1}{4}$ while (5) always holds for $k = 2$. For $k \geq 3$, (5) is equivalent to

$$\alpha \geq \frac{k - 1 - k4^{-1/k}}{k-2}. \tag{6}$$

One can show that $h(x) = (x - 1 - x4^{-1/x})/(x-2)$ is decreasing in x for $x > 2$ and, thus, (6) holds provided that $\alpha \geq h(3) = 2 - \frac{3}{\sqrt[3]{4}}$. □

6. Concluding remarks. The existence of functions with property $\frac{3}{4}$ proves a worst-case bound on the relative duality gap associated with (LP), namely that

$$Z^*_{LP} \leq \frac{4}{3}Z^*_{IP}.$$

Moreover, this worst-case analysis is tight, as can be seen from the MAX 2SAT instance

$$\begin{cases} x_1 \vee x_2 \\ x_1 \vee \overline{x}_2 \\ \overline{x}_1 \vee x_2 \\ \overline{x}_1 \vee \overline{x}_2 \end{cases}$$

with unit weights. As we observed previously, the LP solution $y_i = \frac{1}{2}$ for all i and $z_j = 1$ for all j is optimal for any instance without unit clauses and has value $\sum_j w_j$. In this case, $Z^*_{LP} = 4$, while $Z^*_{IP} = 3$.

The performance guarantee of $\frac{3}{4}$ for our algorithms is also tight. For any instance without unit clauses, all of our algorithms reduce to Johnson's algorithm, since $y_i = \frac{1}{2}$ for all i is an optimal solution to (LP) and all the functions given above have $f(\frac{1}{2}) = \frac{1}{2}$. Johnson's algorithm is a $\frac{3}{4}$-approximation algorithm on this class of instances, and he gives instances of this type that are tight for his algorithm [7]. Furthermore, *any* function f with property $\frac{3}{4}$ must satisfy $f(\frac{1}{2}) = \frac{1}{2}$. This follows from the definition of property $\frac{3}{4}$ for the values $k = 2$, $l = 0$ or 2, and $y_1 = y_2 = \frac{1}{2}$. Therefore, without changing our analysis or strengthening the linear programming relaxation, one cannot expect to beat the performance guarantee of $\frac{3}{4}$.

Results of Arora et al. [2] imply that there exist constants within which MAX 2SAT and MAX 3SAT (every clause has at most 3 literals) cannot be approximated unless $P = NP$. As of the writing of this paper, the best known constant for MAX 3SAT is 112/113 [3]. There is much room for improvement between this hardness result and the approximation algorithms presented here and by Yannakakis [13].

Thus it is an interesting open question as to whether the linear programming relaxation can be strengthened so that a better performance guarantee is possible using these techniques. Recent work of the authors [5] has shown that using a form of randomized rounding on a nonlinear programming relaxation gives a .878-approximation algorithm for the MAX 2SAT problem. It is not yet clear whether this result can be extended to MAX SAT in general. Another interesting open question is that of completely characterizing the functions with property $\frac{3}{4}$. Finally, we would like to know if the technique used here of randomized rounding with a function other than the identity function can be applied to other problems with natural linear programming relaxations.

Acknowledgments. We would like to thank Rick Vohra for suggesting the search for simple functions with property $\frac{3}{4}$.

REFERENCES

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

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, *Proof verification and hardness of approximation problems*, in Proc. 33rd IEEE Symposium on the Foundations of Computer Science, Pittsburgh, PA, 1992, pp. 14–23.

[3] M. Bellare, S. Goldwasser, C. Lund, and A. Russell, *Efficient probabilistically checkable proofs and applications to approximation*, in Proc. 25th ACM Symposium on the Theory of Computing, San Diego, CA, 1993, pp. 294–304.

[4] M. Garey, D. Johnson, and L. Stockmeyer, *Some simplified NP-complete graph problems*, Theoret. Comput. Sci., 1 (1976), pp. 237–267.

[5] M. X. Goemans and D. P. Williamson, *.878-approximation algorithms for MAX CUT and MAX 2SAT.*, in Proc. 26th ACM Symposium on the Theory of Computing, Montreal, 1994, pp. 422–431.

[6] M. X. Goemans and D. P. Williamson, *A new $\frac{3}{4}$-approximation algorithm for MAX SAT*, in Proc. 3rd Integer Programming and Combinatorial Optimization Conference, Erice, Italy, Apr. 1993, pp. 313–321.

[7] D. Johnson, *Approximation algorithms for combinatorial problems*, J. Comput. System Sci., 9 (1974), pp. 256–278.

[8] R. Kohli and R. Krishnamurti, *Average performance of heuristics for satisfiability*, SIAM J. Discrete Math., 2 (1989), pp. 508–523.

[9] K. Lieberherr and E. Specker, *Complexity of partial satisfaction*, J. Assoc. Comput. Mach., 28 (1981), pp. 411–421.

[10] P. Raghavan, *Probabilistic construction of deterministic algorithms: Approximating packing integer programs*, J. Comput. System Sci., 37 (1988), pp. 130–143.

[11] P. Raghavan and C. Thompson, *Randomized rounding: A technique for provably good algorithms and algorithmic proofs*, Combinatorica, 7 (1987), pp. 365–374.

[12] E. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

[13] M. Yannakakis, *On the approximation of maximum satisfiability*, in Proc. 3rd Annual ACM-SIAM Symposium on Discrete Algorithms, Orlando, FL, 1992, pp. 1–9.

SIAM J. DISC. MATH.
Vol. 7, No. 4, pp. 667–679, November 1994

014

HOMOMORPHIC ZERO-KNOWLEDGE THRESHOLD SCHEMES OVER ANY FINITE ABELIAN GROUP*

YVO G. DESMEDT† AND YAIR FRANKEL‡

Abstract. A threshold scheme is an algorithm in which a distributor creates l shares of a secret such that a fixed minimum number (t) of shares are needed to regenerate the secret. A perfect threshold scheme does not reveal anything new from an information theoretical viewpoint to $t-1$ shareholders about the secret. When the entropy of the secret is zero all sharing schemes are perfect, so perfect sharing loses its intuitive meaning. The concept of zero-knowledge sharing scheme is introduced to prove that the distributor does not reveal anything, even from a computational viewpoint. New homomorphic perfect secret threshold schemes over any finite Abelian group for which the group operation and inverses are computable in polynomial time are developed. One of the new threshold schemes also satisfies the zero-knowledge property. A generalization toward a homomorphic zero-knowledge general sharing scheme over any finite Abelian group is discussed and it is proven that ideal homomorphic threshold schemes do not always exist.

Key words. threshold scheme, secret sharing, zero-knowledge, Lenstra constant, algebraic integers

AMS subject classifications. 08A70, 11R04, 20K01, 94A60

1. Introduction.

Secret sharing schemes [5], [29], [17] provide a means to distribute *shares* of a secret so that any subset of individuals (shareholders) specified by an access structure can recompute the secret. Threshold schemes [5], [29] have an access structure where t out of l individuals can recompute the secret. As an example, in Shamir's threshold scheme [29] the distributor generates a random polynomial, $f(x)$, of degree $t-1$ over a finite field such that $f(0)=k$ is the secret and gives each individual i the share $f(i)$. Lagrange interpolation may be used by any t shareholders, at a later stage, to regenerate k. In addition to their use in recomputing secrets, threshold schemes are used, for example, in fault tolerant computing [27] and threshold signatures [9]. Threshold schemes have been studied from many viewpoints such as combinatorics (e.g., [33]) and guarding against cheaters (e.g., [7]). In this paper we study some algebraic and computational aspects of threshold schemes.

Benaloh [1] discussed homomorphic sharing (threshold) schemes as those having the property that when $s_{i,1}$ is i's share of k_1 and $s_{i,2}$ is i's share of k_2, then $s_{i,1}\cdot s_{i,2}$ is i's share of $k_1 * k_2$. For such threshold schemes t shareholders can reconstruct $k_1 * k_2$ using their $s_{i,1}\cdot s_{i,2}$. All of the schemes we will discuss have this property. When both operators "$*$" and "$\cdot$" are identical or when there is no confusion possible, we will simply speak of *multiplicative* threshold schemes (when the operation is addition, we could speak about additive). Homomorphic threshold schemes can be used to set up secret ballot election schemes [1], and existing threshold authentication (and threshold signature) schemes [9] are based on them. To authenticate a message in such

[*] Received by the editors January 13, 1992; accepted for publication (in revised form) November 17, 1993. A part of this research was supported by National Science Foundation grant NCR-9106327. A preliminary version of this paper was presented at Sequences '91 and published in *Sequences* II: *Methods in Communication, Security, and Computer Science*, R. Capocelli, et al., eds., Spinger-Verlag, New York, Berlin, 1993, pp. 369–378.

[†] Department of Electrical Engineering and Computer Science, University of Wisconsin, Milwaukee, P.O. Box 784, Milwaukee, Wisconsin 53201 (desmedt@cs.uwm.edu).

[‡] GTE Laboratories Incorporated, 40 Sylvan Road, Waltham, Massachusetts 02254 (yf01@gte.com). Part of this work was done while this author was at the Department of Electrical Engineering and Computer Science, University of Wisconsin, Milwaukee.

a system, t shareholders out of l use their shares. In such a threshold authentication system no other message can be authenticated without using t shares.

Existing sharing schemes are over fields or over some finite geometries (for an overview on current sharing schemes consult [31]). In this paper we present a method to create a multiplicative (homomorphic) perfect threshold scheme over any finite Abelian group. We require that the group is finite [6].

At first glance it may seem trivial to make a multiplicative secret threshold scheme over a finite Abelian group because of the fundamental theorem of Abelian groups [18] and an appropriate homomorphism [1, p. 255]. However, there are several mathematical and computational problems with this solution (see §4). We solve these problems by requiring only that the group operation and inverses in the group can be calculated in polynomial time.

To guarantee that a threshold scheme is secure Stinson and Vanstone [33] introduced *perfect* threshold schemes. Perfect threshold schemes do not reveal anything new from an information theoretic point of view about the secret when $t-1$ shares are used. We require more. Our scheme does not reveal anything new (about the secret or about whatsoever) even from a computational viewpoint to any $t-1$ shareholders. To formally prove this we introduce a new concept called *zero-knowledge secret sharing.* Perfect zero-knowledge incorporates both aspects. An application of zero-knowledge secret sharing may be found in [9].

In §2 we overview the concept of threshold scheme. Zero-knowledge sharing schemes are defined in §3. The mathematical foundations of this paper are laid in §4. A multiplicative threshold scheme over any finite Abelian group which requires the distributor to know a nontrivial multiple ($\neq 0$) of the exponent of the group is presented in §5. In §6 we present a multiplicative zero-knowledge threshold scheme over finite Abelian groups. Generalizations and optimizations are discussed in §7.

2. Notation. We now introduce formal definitions and notation used in this paper. When $\mathcal{A}$ is a set, $|\mathcal{A}|$ will denote the cardinality of the set; when a is a string, $|a|$ will denote the length of the string; and when r is a real, $|r|$ will denote the absolute value of r.

DEFINITION 2.1. *Let $\mathcal{K}$ be a set with elements called keys, or secrets. A threshold scheme contains two algorithms, one that creates shares of a secret key $k \in \mathcal{K}$ for l individuals so that any t individuals (t is fixed and $t \leq l$) can regenerate the secret using the second algorithm, yet less than t individuals cannot using any method. Let $\mathcal{A} = \{1, \ldots, l\}$ and $\mathcal{S}$ be the set of possible shares. The distributor generates the secret shares $\mathcal{S}_{\mathcal{A}} = (s_1, \ldots, s_l)$ where $s_i \in \mathcal{S}$ and the public directory $\mathcal{X}_{\mathcal{A}} = (x_1, \ldots, x_l)$. For each $\mathcal{B} = \{i_1, \ldots i_{|\mathcal{B}|}\} \subset \mathcal{A}$ with $i_j < i_{j+1}$, we define the selection function $\sigma_{\mathcal{B}} : \mathcal{S}^l \to \mathcal{S}^{|\mathcal{B}|} : (s_1, \ldots, s_l) \mapsto (s_{i_1}, \ldots, s_{i_{|\mathcal{B}|}})$. More formally, a (t, l)-threshold scheme satisfies*

1. *$\forall \mathcal{B} \subset \mathcal{A}$ where $|\mathcal{B}| = t-1$ holds: if $\mathrm{H}(k) \neq 0$ then $0 < \mathrm{H}(k \mid \mathcal{S}_{\mathcal{B}}, \mathcal{X}_{\mathcal{A}}) \leq \mathrm{H}(k)$ for H the entropy function [12] and $\mathcal{S}_{\mathcal{B}} = \sigma_{\mathcal{B}}(\mathcal{S}_{\mathcal{A}})$;*

2. *$\forall \mathcal{B} \subset \mathcal{A}$ where $|\mathcal{B}| = t$, there exists a function $\eta_{\mathcal{B}, \mathcal{X}_{\mathcal{A}}}$ such that $\eta_{\mathcal{B}, \mathcal{X}_{\mathcal{A}}}(\mathcal{S}_{\mathcal{B}}) = k$. Schemes in which for any $\mathcal{B} \subset \mathcal{A}$ with $|\mathcal{B}| = t-1$ holds that $\mathrm{H}(k|\mathcal{S}_{\mathcal{B}}, \mathcal{X}_{\mathcal{A}}) = \mathrm{H}(k)$ are called perfect. When a sharing scheme is perfect and $|\mathcal{K}|/|\mathcal{S}| = 1$, it is called an ideal sharing scheme.*

A more formal definition of the above may be found in [10] and a definition of threshold schemes based on design theory may be found in [33]. We distinguish two phases. The *distribution phase* occurs when a distributor D using a threshold scheme generates $\mathcal{X}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{A}} = (s_1, \ldots, s_l)$ and then publishes $\mathcal{X}_{\mathcal{A}}$ and privately transmits

s_i to each shareholder i. In Shamir's threshold scheme the x_i are different nonzero elements of the finite field identifying i. We call the phase in which k is recomputed using $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}$ the *recomputation phase.*

DEFINITION 2.2. *Let $\mathcal{K}_n$ be a family of finite Abelian groups indexed by $n \in \mathcal{J}$, where $\mathcal{J}$ is an infinite subset of $\{0,1\}^*$. We say that a family of groups $\mathcal{K}_n$ is polynomial time when* (1) *the group operation and inverses may be computed in polynomial time in function of* $|n|$; (2) *generating uniform random elements of $\mathcal{K}_n$ can be performed in probabilistic polynomial time*; *and* (3) *it is easy to check membership in $\mathcal{J}$* ($\mathcal{J} \in$ **BPP** [13]). *For a polynomial time family of groups, $\mathcal{K}_n$, we say that a family of (t,l)-threshold schemes is practical when their distribution phase and recomputation phase can be computed respectively in polynomial time in function of* $\max(l, |n|)$ *and* $\max(t, |l|, |n|)$.

From now on we say $\mathcal{K}_n$ instead of a family of Abelian groups $\mathcal{K}_n$ and we say a threshold scheme over $\mathcal{K}_n$ instead of a family of threshold schemes over $\mathcal{K}_n$. Without loss of generality, we assume that $\mathcal{K}(*)$ is a multiplicative Abelian group. We will usually write $a * b$ as ab when the context is clear. Let $e_\mathcal{K}$ denote the exponent of $\mathcal{K}$ (i.e., the smallest positive integer such that $x^{e_\mathcal{K}} = 1$ for all x in $\mathcal{K}$). We now overview the definition of homomorphic threshold schemes [1].

DEFINITION 2.3. *If we have the operation "$\cdot$" over $\mathcal{S}$ (so $\mathcal{S}$ is closed under "$\cdot$"), and $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}$ is a homomorphism from $\mathcal{S}^t(\cdot)$ to $\mathcal{K}(*)$ for all $\mathcal{B} \subset \mathcal{A}$ with $|\mathcal{B}| = t$, then the threshold scheme is called a homomorphic threshold scheme.*[1] *When both operations are identical we will just speak about multiplicative threshold scheme (or additive when appropriate).*

3. Zero-knowledge sharing scheme.

3.1. Informal discussion. When the entropy of a secret is zero all sharing schemes are perfect, so perfect sharing loses its intuitive meaning. Notice that the secret key in a public key system has entropy zero. So a sharing scheme that gives the secret key when its entropy is zero to all shareholders would satisfy the requirement of being perfect. However, it is clear that this is not secure. This is not the intent of threshold schemes. We now discuss how to solve this problem.

First observe that in a threshold scheme a key distributor gives l shareholders *secret shares* s_i and *public information* x'_i that allow any t of the shareholders to recalculate the secret. So it is clearly desirable that $t-1$ shares will not enable those shareholders to *calculate* anything new about the secret. However, the public information x'_i could reveal something too—for example, the exponent $e_\mathcal{K}$ of the group—but $e_\mathcal{K}$ must be kept secret in many cryptographic algorithms [25], [28]. To formalize this we introduce zero-knowledge threshold schemes which broaden the concept of perfect threshold schemes. Informally a zero-knowledge threshold scheme satisfies the property that the key distributor is *not* giving anything new from a computational viewpoint to any subset of $t-1$ shareholders. Observe that the definition of perfect threshold scheme informally says that the key distributor does not give, to any subset of $t-1$ shareholders, anything new about the secret from an *information theoretical viewpoint*. Our concept implies that nothing is revealed to $t-1$ shareholders about anything. Perfect zero-knowledge encompasses the information theoretical and computational viewpoints.

Because the key k could reveal something about the group that t shareholders could not calculate, we used $t-1$ in the above discussion. Indeed when the key k is a

[1] If the threshold scheme is not perfect then the above definition must be slightly adapted.

nontrivial square root of 4 modulo $n = p \cdot q$, p and q being primes, then t shareholders can factor n.

Informally a threshold scheme is perfect zero-knowledge when $t-1$ shares and the public information can be simulated by anyone who does not know the secret, such that the simulated information and the real one occur with the same probability.

To guarantee that t shareholders (or more) do not learn anything more than what follows from their joint knowledge of k, we require that the protocol is minimal knowledge [11]. Informally a threshold scheme is perfectly minimal knowledge when t shares and the public information can be simulated given the secret key k.

The concepts of zero-knowledge and minimal-knowledge *threshold* schemes can easily be broadened toward sharing schemes with other access structures.

We now give a formal definition. Readers who understand the above informal discussion about zero-knowledge threshold scheme can skip the formal definition below.

3.2. Formal definition. Let us first adapt the definition of zero-knowledge [14] to our needs. We have to take into consideration that when r is exponential in length of the input, r machines have more power than a single one.

DEFINITION 3.1. *For a language (a set) $L \subset \{0,1\}^*$ and $x \in L$, two families of random variables $\{U(x)\}$ and $\{V(x)\}$ are equal on L when $\{U(x)\} = \{V(x)\}$ for all $x \in L$. These families of random variables are r-parallel statistically indistinguishable on L if, for all constants $c > 0$ and all sufficiently long $x \in L$,*

$$\sum_{\alpha \in \{0,1\}^*} |\operatorname{prob}(U(x) = \alpha) - \operatorname{prob}(V(x) = \alpha)| < r^{-1} \cdot |x|^{-c}.$$

When appropriate a probabilistic Turing machine D' of which the expected computation time is bounded by $r|x|^c$, with c a constant, will be called a simulator. For a protocol $(D, \mathcal{B})$ where $|\mathcal{B}| = r$, the Joint-view of $\mathcal{B}$ is what $\mathcal{B}$ sees (i.e., the concatenated strings received from D, and when the machines in $\mathcal{B}$ use randomness, the string is concatenated with these random strings) and the Joint-view$_{D,\mathcal{B}}(x,h)$ is the random variable whose value is their view (i.e., $\mathcal{B}$'s view) for input x and history tape h.

History tapes allow the parties to be involved in many consecutive protocols. Therefore, they are a crucial part in the definition of zero-knowledge [14].

DEFINITION 3.2. *Let the Joint-view of $\mathcal{B}$ for a fixed $k \in \mathcal{K}$ be $(\mathcal{S}_{\mathcal{B}}, \mathcal{X}_{\mathcal{A}})_k$. Let D_k be the distributor of the key $k \in \mathcal{K}$. Let $\mathcal{B}'$ be a set of possibly dishonest shareholders such that $|\mathcal{B}'| \leq t-1$. Let D' be a simulator. A threshold scheme is called perfect (statistical) zero-knowledge when: $\forall \mathcal{B}' : \exists D' : \forall k \in \mathcal{K} : \{D'(x,h)\}$ and $\{$Joint-view$_{D_k,\mathcal{B}'}(x,h)\}$ are equal ($|\mathcal{B}'|$-parallel statistically indistinguishable) on $L' = \{(x,h) \mid x \in L$ and $h \in \{0,1\}^*$ and $|h| \leq |\mathcal{B}'| \cdot |x|^c$ for c a constant$\}$ where $L = \{(n,t,l) \mid t,l \in Z^+$, $0 < t \leq l$ and $n \in \mathcal{J}\}$.*

Observe that our definition is very strong, indeed a weak (less secure) definition would allow D' to run in $t^d|n|^c$ for c and d constants. The simulator D' does not know k. The order of the quantifiers is therefore crucial. Computational zero-knowledge sharing schemes can be defined similarly using [14].

DEFINITION 3.3. *Let $\mathcal{B}$ be a set of shareholders such that $|\mathcal{B}| \geq t$. Let D' be a simulator who has a one-time access to an oracle that gives D' the secret key k. A threshold scheme is called perfect (statistical) minimal-knowledge when: $\forall \mathcal{B} : \exists D' : \forall k \in \mathcal{K} : \{D'(x,h)\}$ and $\{$Joint-view$_{D_k,\mathcal{B}}(x,h)\}$ are equal ($|\mathcal{B}|$-parallel statistically indistinguishable) on $L' = \{(x,h) \mid x \in L$ and $h \in \{0,1\}^*$ and $|h| \leq |\mathcal{B}| \cdot |x|^c$ for c a constant$\}$ where $L = \{(n,t,l) \mid t,l \in Z^+$, $0 < t \leq l$ and $n \in \mathcal{J}\}$.*

4. Foundation.

4.1. First approach. Let $\{g_1, \ldots, g_h\}$ be a set of generators for $\mathcal{K}$. Using [29], the distributor chooses independent uniformly random polynomials $f_j(x)$ for $1 \leq j \leq h$ of degree $t-1$ such that the secret $k = g_1^{\gamma_1} \ldots g_h^{\gamma_h} \in \mathcal{K}$ with $\gamma_j = f_j(0)$. Let the share for individual i be $s_i = g_1^{f_1(i)} \ldots g_h^{f_h(i)} \in \mathcal{K}$. Then if $e_{\mathcal{K}}$ is known to the shareholders, any subgroup of t shareholders can easily recompute k using Lagrange interpolation. However, there are mathematical and computational problems with this solution [1, p. 255]. First, this scheme is useless when $e_{\mathcal{K}} = 2$ since l can only be 1. Second, the distributor must solve the *discrete logarithm problem*, which is considered hard [8], [22], [24] for many groups even when generators are known. Third, a multiple of $e_{\mathcal{K}}$ must be known by the distributor and shareholders to use Lagrange interpolation, but this reduces the usefulness of this scheme since multiples of $e_{\mathcal{K}}$ must be kept secret for many cryptographic algorithms, e.g., the order of Z_n^* [25], [28]. Finally, the distributor must know a set of generators.

4.2. Basic solution. To solve the above first problem (i.e., $e_{\mathcal{K}} = 2$) our schemes work over $\mathcal{K}^d = \mathcal{K} \times \cdots \times \mathcal{K}$, the direct product of $\mathcal{K}'s$. We initially assume that the exponent of $\mathcal{K}$ is a prime q and treat the Abelian group $\mathcal{K}^d$ as a vector space over $GF(q^d)$. This will be modified to a module over $Z[u]$, an algebraic extension of Z. We consider $\mathcal{K}$ as a multiplicative group, but denote the operation in $\mathcal{K}^d$ as vector addition.

Let d be an integer such that $q^d - 1 \geq l$. Let u be the root to a monic irreducible polynomial $p(z)$ of degree d used to define $GF(q^d)$, so $u^d = a_0 + \cdots + a_{d-1}u^{d-1}$. Elements of the vector space $\mathcal{K}^d$ are written as $\vec{k} = [k_0, \ldots, k_{d-1}]$ and the identity element is $\vec{0} = [1, \ldots, 1]$. We remark that this vector space is *not* necessarily of dimension d and the notation $[k_0, \ldots, k_{d-1}]$ should not be confused with coordinates. Addition in $\mathcal{K}^d$ is defined as $[k_0, \ldots, k_{d-1}] + [k'_0, \ldots, k'_{d-1}] = [k_0 * k'_0, \ldots, k_{d-1} * k'_{d-1}]$. For $(b_0 + \cdots + b_{d-1}u^{d-1}) \in GF(q^d)$ and $[k_0, \ldots, k_{d-1}] \in \mathcal{K}^d$ the scalar operation "$\cdot$" for the vector space is defined as $(b_0 + \cdots + b_{d-1}u^{d-1}) \cdot [k_0, \ldots, k_{d-1}] = \sum_{i=0}^{d-1} b_i u^i [k_0, \ldots, k_{d-1}]$, where $b \cdot [k_0, \ldots, k_{d-1}] = [k_0^b, \ldots, k_{d-1}^b]$ for $b \in GF(q)$ and $u \cdot [k_0, k_1, \ldots, k_{d-1}] = [1, k_0, \ldots, k_{d-2}] + [k_{d-1}^{a_0}, k_{d-1}^{a_1}, \ldots, k_{d-1}^{a_{d-1}}]$, recursively $bu^i \cdot [k_0, \ldots, k_{d-1}] = u \cdot (u^{i-1} \cdot (b \cdot [k_0, \ldots, k_{d-1}]))$. We leave it to the reader to prove that this is a vector space.[2]

5. Elementary approach.

Only in this section we assume that a nontrivial multiple ($\neq 0$) of the exponent, $e'_{\mathcal{K}}$, is known to the distributor.

5.1. Exponent is a prime. We now discuss a sharing scheme over an Abelian group where the exponent is a prime q.

Scheme 1.

Distribution phase. The distributor D chooses $p(z)$ as specified in §4 using [4], [26]. Then D chooses l different elements $x_i \in GF(q^d) \setminus 0$ and $t-1$ independent uniformly random elements $\vec{s}_i \in \mathcal{K}^d$ $(1 \leq i \leq t-1)$. Let k_0 be the secret in $\mathcal{K}$ and $\vec{k} = [k_0, 1, \ldots, 1]$ represent the secret in $\mathcal{K}^d$. Then D computes

$$\vec{s}_j = y_{j,C_j}^{-1} \cdot \left(\vec{k} - \left(\sum_{\substack{i \neq j \\ i \in C_j}} y_{i,C_j} \cdot \vec{s}_i \right) \right) \tag{1}$$

[2] So $\mathcal{K}^d$ is the tensor product of modules: $GF(q^d) \otimes_{GF(q)} \mathcal{K}$.

for $t \leq j \leq l$ where $\mathcal{C}_j = \{1, 2, \ldots, t-1, j\}$ and

$$y_{i,\mathcal{B}} = \prod_{\substack{j \notin \mathcal{B} \\ j \in \mathcal{A}}} (x_i - x_j) \prod_{\substack{j \in \mathcal{B} \\ j \neq i}} (0 - x_j). \tag{2}$$

D publishes $x'_i = (x_i, p(z))$ for all $i \in \mathcal{A}$ and privately transmits to shareholder i the share $\vec{s}_i = [s_{i,0}, \ldots, s_{i,d-1}]$.

Recomputation phase. The t shareholders in $\mathcal{B}$ compute $y_{i,\mathcal{B}}$ in $Z[u]$ and then calculate $k_0 = F_0(\vec{k}) = \prod_{i \in \mathcal{B}} F_0(y_{i,\mathcal{B}} \cdot \vec{s}_i)$ where $F_0 : \mathcal{K}^d \to \mathcal{K} : [k'_0, \ldots, k'_{d-1}] \mapsto k'_0$.

Observe that the shareholders do not need to know q.

LEMMA 5.1. *For any finite Abelian group where $e_{\mathcal{K}}$ is a prime q, Scheme 1 is a multiplicative perfect (t, l)-threshold scheme.*

Proof. We will prove a more general result, i.e., that our scheme can be adapted to an ideal sharing scheme over $\mathcal{K}^d$. The lemma then follows trivially.

Let $\{g_1, \ldots, g_h\}$ be an independent set of generators for $\mathcal{K}$. Due to the fundamental theorem of Abelian groups, h is invariant. We first show that $\mathcal{K}^d$ is a vector space of dimension h. For $\vec{g}_i = [g_i, 1, \ldots, 1] \in \mathcal{K}^d$ for $1 \leq i \leq h$, note that $(b_0 + \cdots + b_{d-1}u^{d-1}) \cdot \vec{g}_i = [g_i^{b_0}, \ldots, g_i^{b_{d-1}}]$. Thus $\{\vec{g}_1, \ldots, \vec{g}_h\}$ is a basis for the vector space $\mathcal{K}^d$. Note that the $y_{i,\mathcal{B}}$ is a unit since we are working in $GF(q^d)$ and $x_i = x_j$ if and only if $i = j$.

First observe that Shamir's threshold scheme over a finite field can be modified. Instead of giving shareholder i the share $f(x_i)$, one can give as share $\alpha'_i = f(x_i) / \prod_{\substack{j \neq i \\ j \in \mathcal{A}}} (x_i - x_j)$. The key may then be computed by performing $f(0) = \sum_{i \in \mathcal{B}} y_{i,\mathcal{B}} \cdot \alpha'_i$ with $y_{i,\mathcal{B}}$ as in (2). Second, it is easy to prove that Shamir's scheme is also an ideal threshold scheme over any module provided x_i and $x_i - x_j$ are units.

All $\vec{s}_i$ for $1 \leq i \leq l$ can be written as $\alpha_{1,i} \cdot \vec{g}_1 + \cdots + \alpha_{h,i} \cdot \vec{g}_h$ and $\vec{k} = \gamma_1 \cdot \vec{g}_1 + \cdots + \gamma_h \cdot \vec{g}_h$. For all m $(1 \leq m \leq h)$ it holds that γ_m and $\alpha_{m,i}$ for $1 \leq i \leq t-1$ define a polynomial $f_m(x)$ of degree $t-1$ over $GF(q^d)$ such that $f_m(0) = \gamma_m$. To prove that our scheme is a threshold scheme, we need only to remark that for all $i : \alpha_{m,i} = f_m(x_i) / \prod_{\substack{j \neq i \\ j \in \mathcal{A}}} (x_i - x_j)$ due to our first observation and that we have the isomorphism $\mathcal{K}^d(+) \to GF(q^d) \times \cdots \times GF(q^d) \,:\, \vec{k} = \gamma_1 \cdot \vec{g}_1 + \cdots + \gamma_h \cdot \vec{g}_h \mapsto (\gamma_1, \ldots, \gamma_h)$.

To demonstrate that the scheme is perfect it is sufficient to prove that $\text{prob}(\vec{\mathbf{k}} = \vec{k} | \boldsymbol{\mathcal{S}}_{\mathcal{B}'} = \mathcal{S}_{\mathcal{B}'}) = \text{prob}(\vec{\mathbf{k}} = \vec{k})$ for any $\mathcal{B}'$ such that $|\mathcal{B}'| = t-1$. This can be proven by showing that for any given $\vec{k}' \in \mathcal{K}^d$ and $\mathcal{S}_{\mathcal{B}'}$ there exists a unique element $\vec{s}_j \in \mathcal{K}^d$ such that $\vec{k}' = \sum_{i \in \mathcal{B}} y_{i,\mathcal{B}} \cdot \vec{s}_i$ where $\mathcal{B} = \mathcal{B}' \cup \{j\}$. Indeed $y_{j,\mathcal{B}}$ are invertible, so $\vec{s}_j = y_{j,\mathcal{B}}^{-1} \cdot (\vec{k}' - (\sum_{\substack{i \neq j \\ i \in \mathcal{B}}} y_{i,\mathcal{B}} \cdot \vec{s}_i))$. □

We say a vector space $\mathcal{K}_n$ over $\mathcal{F}_n$ is polynomial time if its additive group is polynomial time, if the scalar operation can be performed in polynomial time and when generating uniform random elements of $\mathcal{K}_n$ can be performed in probabilistic polynomial time.

COROLLARY 5.2. *For any polynomial time finite vector space $\mathcal{K}_n$ over a finite field $\mathcal{F}_n$, Scheme 1 can be adapted to a practical additive ideal (t, l)-threshold scheme when $l < |\mathcal{F}|$ and to a perfect one otherwise.*

Although many of the earlier proposed sharing schemes over finite geometries [31] can trivially be adapted to vector spaces, not all are practical for any polynomial time vector space. Indeed using §4, it is not difficult to make vector spaces for which it is believed to be hard to find coordinates.

5.2. General case. Let $e'_{\mathcal{K}} = q_1^{\omega_1} \cdots q_c^{\omega_c}$ be a multiple of $e_{\mathcal{K}}$ where q_m is a prime ($1 \leq m \leq c$), $q_m < q_{m+1}$ and ω_m are integers > 0, and all q_m are known to the distributor.

Scheme 2.

Distribution phase. Let d be an integer such that $q_1^d - 1 \geq l$. The distributor D chooses monic irreducible polynomials $p_m(z)$ of degree d over Z_{q_m} for all $m \leq c$ [4], [26]. Using the Chinese remainder theorem on the *coefficients* of the polynomial, D creates a new polynomial $p(z) \in Z_{e'_{\mathcal{K}}}[z]$ with root u, so that $p(z) \equiv p_m(z) \bmod q_m^{\omega_m}$ for all m. For each $i \in \mathcal{A}$, D chooses a unique element $w_{i,m} = b_{i,m,0} + \cdots + b_{i,m,d-1}u^{d-1} \in Z^0_{q_m}[u]$ and then computes $x_i \in Z_{e'_{\mathcal{K}}}[u]$ where $x_i \equiv w_{i,m} \bmod q_m^{\omega_m}$ using the Chinese remainder theorem on the coefficients of $w_{i,m}$ for each $m \leq c$. To compute the $\vec{s}_i$, the distributor proceeds as before except that $\vec{s}_i$ are elements[3] of the module $\mathcal{K}^d$ over $Z_{e'_{\mathcal{K}}}[u]$.

Recomputation phase. The recomputation phase is similar to the recomputation in Scheme 1, but working in a module over $Z[u]$.

THEOREM 5.3. *For any finite Abelian group $\mathcal{K}$ and any l, Scheme 2 is a multiplicative perfect (t, l)-threshold scheme and when l is less than the smallest prime factor of $e'_{\mathcal{K}}$ Scheme 2 is an ideal homomorphic threshold scheme.*

Proof. Observe that all that is needed is a polynomial $p(z)$ and x_i for each $i \in \mathcal{A}$ such that x_i and $(x_i - x_j)$ are relatively prime to $p(z)$ in Z_{q_m} for all $i, j \in \mathcal{A}$ and m such that $i \neq j$ and $1 \leq m \leq c$. The distribution phase of Scheme 2 constructs such a $p(z)$ and x_i. A similar proof as in Lemma 5.1 is used to prove that this is a perfect threshold scheme.

When l is less than the smallest prime q_1, then the scheme can work in $Z_{e'_{\mathcal{K}}}$ since $x_i - x_j$ has an inverse when $1 \leq x_i, x_j \leq l$ and $x_i \neq x_j$, making it ideal for this case. □

5.3. A security problem. The disadvantage with Schemes 1 and 2 is that the polynomial $p(z)$ can reveal information about $e_{\mathcal{K}}$ to the shareholders. Indeed if $p(z)$ is not irreducible modulo a prime q then the shareholders will know that q does not divide $e_{\mathcal{K}}$. Moreover to execute the algorithm in Theorem 5.3, the distributor needs to know the factorization of $e'_{\mathcal{K}}$ (although by modifying above, knowing $e'_{\mathcal{K}}$ is sufficient). In the next section these problems will be resolved.

6. Zero-knowledge version. In this section we do not assume that a multiple of the exponent, $e'_{\mathcal{K}}$, is known to the distributor. We now propose a multiplicative perfect zero-knowledge (t, l)-threshold scheme for any family of finite Abelian groups $\mathcal{K}_n$ and any reasonably large l without requiring that the exponent be known to anyone. The concept of zero-knowledge threshold scheme will be used to prove that $t - 1$ shareholders will not obtain any additional information (about the group) from the distributor.

Scheme 3.

Distribution phase. Let q be a prime greater than or equal to $l + 1$. Let u be a root[4] of the cyclotomic polynomial $p(z) = \sum_{j=0}^{q-1} z^j$ and $x_i = \sum_{j=0}^{i-1} u^j$. Let k_0 be the

[3] We abuse the vector notation to make the analogy with the vector space case clear.

[4] Formally, $Z[u]$ is isomorphic to $Z[z]/(p(z))$.

secret in $\mathcal{K}$ and $\vec{k} = [k_0, 1, \ldots, 1]$ represent the secret in $\mathcal{K}^{q-1}$. Then D chooses $t-1$ independent uniformly random elements $\vec{s}_j \in \mathcal{K}^{q-1}$ $(1 \le j \le t-1)$ and computes $\vec{s}_j$ as in (1) for $t \le j \le l$ except

$$y_{i,\mathcal{B}} = \prod_{\substack{j \in \mathcal{B} \\ j \neq i}} (x_i - x_j)^{-1}(0 - x_j) \in Z[u] \tag{3}$$

and the computation for $\vec{s}_j$ is performed in the module $\mathcal{K}^{q-1}$ over $Z[u]$. The distributor D privately transmits to shareholder i the share $\vec{s}_i = [s_{i,0}, \ldots, s_{i,q-2}]$.

Recomputation phase. The t shareholders in $\mathcal{B}$ compute $y_{i,\mathcal{B}}$ as in (3) in $Z[u]$ and then calculate $k_0 = F_0(\vec{k}) = \prod_{i \in \mathcal{B}} F_0(y_{i,\mathcal{B}} \cdot \vec{s}_i)$. See Corollaries 6.2 and 6.3 for details about the algorithm.

THEOREM 6.1. *Scheme 3 is a multiplicative perfect zero-knowledge, miminal-knowledge (t,l)-threshold scheme for any polynomial time finite Abelian group $\mathcal{K}_n$ provided that $l = O(|n|^c)$ for c a constant.*[5]

Proof. We note that the $\gcd(z^m - 1, z^n - 1) = z^d - 1$ for $d = \gcd(m,n)$ over any field [3, p. 29]. Thus over any field $\gcd(p(z), x_i) = \gcd((z^q - 1)/(z-1), (z^i - 1)/(z-1)) = 1$ for a prime q and $1 \le i < q$. Moreover, $\gcd(p(z), z^i) = 1$, thus $y_{i,\mathcal{B}}$ is invertible in $Z[u]$ (see also [15]) and can be calculated in polynomial time without having any knowledge of $e_{\mathcal{K}}$ or a multiple of it. The rest follows from Theorem 5.3.

It is obvious that this is a perfect zero-knowledge, minimal-knowledge threshold scheme. Any $t-1$ shares $\vec{s}_i$ $(i \in \mathcal{B} : |\mathcal{B}| = t-1)$ occur with uniform probability as proved in Lemma 5.1 and x_i $(i \in \mathcal{A})$ are a polynomial time deterministic function of i. So a simulator D' can generate $t-1$ of such $\vec{s}_i$ and all x_i with the same distribution.

Finally, due to Bertrand's Postulate [16] (first proved by Tschebyschef) there exists a prime q between l and $2l$. □

COROLLARY 6.2. *In the zero-knowledge threshold scheme each shareholder in the recomputation phase performs $O(tl^2)$ group operations and $O(l)$ inverses. In addition to the time needed to choose $t-1$ shares, the distribution phase takes $O(t^2l^2(l-t))$ group operations and $O(tl(l-t))$ inverse operations. So when $l = O(|n|^c)$, with c a constant, the threshold scheme is practical.*

Proof. Using Horner's rule $(b_i u^i + b_{i-1}u^{i-1} + \cdots + b_0)\vec{s}_j = (\cdots u(u(b_i\vec{s}_j) + b_{i-1}\vec{s}_j) + \cdots + b_0\vec{s}_j)$ takes $(2i + \sum_{h=0}^{i+1} |b_h|)(q-1)$ group operations where $|b_h| = \lceil \log_2 b_h \rceil$. Observe that $((-1)^{v_1}a(u)(-1)^{v_2}b(u)) \cdot \vec{k}' = (-1)^{v_1+v_2}a(u)(b(u) \cdot \vec{k}')$ for $a(z)$ and $b(z)$ polynomials. When $i > j$, the elements $(x_i - x_j)^{-1}$ are of the form $(u^j(1 + u + \cdots + u^{i-j-1}))^{-1}$ and by Lemma A.1 in the Appendix $(1 + u + \cdots + u^{i-j-1})^{-1}$ is of the form $\sum_{i=0}^{q-2} b_i u^i$ where $b_i \in \{-1, 0\}$ or $b_i \in \{0, 1\}$. This time the shareholder does not compute $y_{i,\mathcal{B}}$ in $Z[u]$ and $y_{i,\mathcal{B}} \cdot \vec{s}_i$. Now the shareholder will compute $(x_i - x_{j_t})^{-1}(\cdots((x_i - x_{j_1})^{-1}((0 - x_{j_t})(\cdots((0 - x_{j_1}) \cdot \vec{s}_i))) \cdots) \cdots)$ where $j_1, \ldots, j_t \in \mathcal{B} \setminus \{i\}$ taking our previous remarks into consideration. Therefore it takes $O(tl^2)$ group operations and zero or $q-1$ inverses for each shareholder to compute $y_{i,\mathcal{B}} \cdot \vec{s}_i$ and $O(t)$ group operations to compute $k_0 = F_0(\vec{k}) = \prod_{i \in \mathcal{B}} F_0(y_{i,\mathcal{B}} \cdot \vec{s}_i)$ given $y_{i,\mathcal{B}} \cdot \vec{s}_i$ for all $i \in \mathcal{B}$. The distributor takes $O(t^2l^2(l-t))$ group operations and $O(tl(l-t))$ inverse operations to compute all the shares. □

[5] A weaker form of zero-knowledge could allow the simulator D' to run in $t^d|n|^c$, for c and d constants. Using this weaker definition l could be $O(t^{d-1}|n|^e)$, e a constant.

When the distributor for $i \in \{1, 2, \cdots, t-1\}$ splits up the computation of $y_{i,\mathcal{C}_j}$ as $y_{i,\mathcal{C}_j} = y'_i \cdot y''_{i,\mathcal{C}_j}$ where y'_i is independent of j, the distribution phase can be sped up. Hereto he precomputes $y'_i \cdot \vec{s}_i$.

A different method may be used when the group operation is slower than integer multiplications.

COROLLARY 6.3. *In the zero-knowledge threshold scheme each shareholder in the recomputation phase performs $O(tl \log_2 l)$ group operations, $O(l)$ inverses, and $O(t^3 l^2$ $(\log_2 l)^2)$ elementary integer multiplications. In addition to the time needed to choose $t-1$ shares, the distributor performs $O(t^2 l(l-t) \log_2 l)$ group operations, $O(tl(l-t))$ inverse operations, and $O(t^4 l^2 (l-t)(\log_2 l)^2)$ elementary integer multiplications.*

Proof. The shareholder will first compute $y_{i,\mathcal{B}}$ in $Z[u]$. Let $p(z) = \sum_{j=0}^{q-1} z^j$ be the cyclotomic polynomial used in Scheme 3 and u be its root. Let $A_i(u) = \sum_{j=0}^{q-2} a_{i,j} u^j$ where $a_{i,j} \in \{-1, 0\}$ or $a_{i,j} \in \{0, 1\}$. Note that $B(u) = A_1(u) A_2(u) = \sum_{i=0}^{2q-4} b_i u^i$ where $b_i = \sum_{j=0}^{i} a_{1,j} \cdot a_{2,i-j}$. Observe that $u^{q-1} = -\sum_{j=0}^{q-2} u^j$ and $u^i = u^j \bmod p(u)$ where $i \equiv j \bmod q$. Thus the absolute value of the largest coefficient in $B(u) \bmod p(u)$ is $3(q-1)$. Inductively, the largest coefficient of $B(u) = \prod_{j=1}^{h} A_j(u)$ in absolute value is $3^{h-1}(q-1)^{h-1}$. Therefore the number of bits required to represent the largest coefficient in absolute value in $y_{i,B}$ is in $O(t \log_2 l)$. There are $O(tl^2)$ integer multiplications giving $O(t^3 l^2 (\log_2 l)^2)$ elementary integer multiplications (FFT speeds this up).

Let $\vec{s} = [s_0, \ldots, s_{q-2}]$, $s_{q-1} = 1$ and the indices be calculated mod q. Then by induction $F_0(u^j \cdot \vec{s}) = F_0([s_{q-1-j}^{-1} \cdot s_{q-j}, s_{q-1-j}^{-1} \cdot s_{q-j+1}, \ldots, s_{q-1-j}^{-1} \cdot s_{q-j+q-2}]) = s_{q-1-j}^{-1} \cdot s_{q-j}$ because $u^{q-1} = \sum_{j=0}^{q-2} -u^j$. When $y_{i,\mathcal{B}} = y_{i,\mathcal{B},0} + y_{i,\mathcal{B},1} u + \ldots + y_{i,\mathcal{B},q-2} u^{q-2}$ where $y_{i,\mathcal{B},j} \in Z$, then $F_0(y_{i,\mathcal{B}} \cdot \vec{s}_i)$ is calculated as $\prod_{j=0}^{q-2} (F_0(u^j \vec{s}_i))^{y_{i,\mathcal{B},j}}$. Thus each shareholder performs $O(lt \log_2 l)$ group operations and $O(l)$ inverses. The distributor takes $O(t^2 l(l-t) \log_2 l)$ group operations and $O(tl(l-t))$ inverse operations to compute all the shares. □

7. Generalization and optimization.

7.1. Perfect general sharing over a group. General sharing schemes with an access structure different from that of a threshold scheme were introduced in [17]. Monotone access structures satisfy the property that when a set of shareholders $\mathcal{B}$ can recompute a secret then any superset $\mathcal{B}' \supset \mathcal{B}$ can also recompute the secret. We now present a multiplicative general sharing scheme over any finite Abelian group.

Using finite projective geometry, a method to create sharing schemes for any monotone access structure has been developed [32]. Because our next approach is an adaptation of their scheme, let us briefly review their scheme. In their scheme a public hyperplane V_d intersects with a secret hyperplane V_i at a point that is the secret. Points are given to each shareholder in such a way that they meet the following two conditions. First, when a set of shareholders allowed by the access structure work together, they will be able to generate the secret hyperplane V_i. Second, when a set of shareholders not allowed by the access structure work together, they do not obtain any information about the secret point (other than it is contained in V_d).

LEMMA 7.1. *There exists a perfect multiplicative sharing scheme over any finite Abelian group.*

Proof. We modify the scheme developed in [32]. When in [32] the distributor gives a point p_i to shareholders $\{j_1, \ldots, j_{h_i}\}$, the distributor here will give $s_i \in \mathcal{K}$ to shareholders $\{j_1, \ldots, j_{h_i}\}$. Let the total number of such points p_i in [32] be m, then $\prod_{1 \le i \le m} s_i = k$ where $s_1, \ldots, s_{m-1}$ have been chosen as independently uniformly random elements in $\mathcal{K}$. The fact that the sharing scheme is perfect follows from the one-time-pad [30]. □

Note that the scheme is also a sharing scheme over any finite group, but then it is not necessarily homomorphic. The scheme is zero-knowledge if for each set of shareholders $\mathcal{B}'$, which is not allowed by the access structure to recompute k, the total number of s_i that $\mathcal{B}'$ have is $O(|\mathcal{B}'||n|^c)$ (this condition is sufficient, but it is not necessary). Unfortunately, the scheme in [32] generates an exponential number of points for many access structures (including those of some threshold schemes). For those, clearly the scheme in the proof of Lemma 7.1 does not satisfy the zero-knowledge condition.

7.2. Ideal homomorphic threshold schemes. Since the zero-knowledge threshold scheme in §6 is not ideal, one wonders if it is possible to make ideal homomorphic threshold schemes over any (family of) finite Abelian groups. We now answer this question.

THEOREM 7.2. *There exists an infinite number of Abelian groups $\mathcal{K}$ for which there does* not *exist an ideal homomorphic threshold scheme when $l > 2$, even when $l < |\mathcal{K}|$ and $t = 2$.*

Proof. Let $\mathcal{K}$ be finite. We first prove that in any ideal (i.e., $|\mathcal{S}| = |\mathcal{K}|$) threshold scheme when $s_{i_1}, \ldots, s_{i_t} \in \mathcal{S}$ then $S_\mathcal{B} = (s_{i_1}, \ldots, s_{i_t})$ is a valid tuple of shares. Clearly, if $S'_\mathcal{B} = (s'_{i_1}, \ldots, s'_{i_t})$ is a valid tuple of shares then $S''_\mathcal{B} = (s_{i_1}, s'_{i_2}, \ldots, s'_{i_t})$ is also for the scheme to be perfect. Repeating this process for $i_2, \ldots, i_t$ proves that *any* combination of t elements of $\mathcal{S}$ can be valid tuple of t shares.

We now prove that if a homomorphic threshold scheme is ideal then the set $\mathcal{S}$ is a group that is isomorphic to $\mathcal{K}$. Let $s \in \mathcal{S}$ and $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(s, \ldots, s) = k \in \mathcal{K}$. For the threshold scheme to be perfect one needs that $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(s, \ldots, s, \mathcal{S}) = \mathcal{K}$. First $s \cdot \mathcal{S} = \mathcal{S}$ since $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(s, \ldots, s) * \eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(s, \ldots, s, \mathcal{S}) = \eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(s \cdot s, \ldots, s \cdot s, s \cdot \mathcal{S}) = k * \mathcal{K} = \mathcal{K}$. Since $s \cdot \mathcal{S} = \mathcal{S}$ and $\mathcal{S}$ is finite, there exists an element e_x for every $x \in \mathcal{S}$ such that $x \cdot e_x = x$. From this we note that $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(e_x, \ldots, e_x) = 1 \in \mathcal{K}$ since $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(x, \ldots, x) = \eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(x \cdot e_x, \ldots, x \cdot e_x)$. Now, $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(x, \ldots, x, y) = \eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(x \cdot e_x, \ldots, x \cdot e_x, y \cdot e_x) = \eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(x, \ldots, x, y \cdot e_x)$. Since $|\mathcal{S}| = |\mathcal{K}|$ and the scheme is perfect, $y \cdot e_x = y$. Thus e_x is a right identity element. Similarly we can prove a left identity element. So there exists an identity element $1 \in \mathcal{S}$. Let $\psi_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}(x) = \eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(1, \ldots, 1, x)$ where x is the ith share. The mapping $\psi_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}$ is a homomorphism from $\mathcal{S}$ to $\mathcal{K}$. Observe that $\psi_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}$ is an onto mapping because the scheme is perfect. The fact that $|\mathcal{S}| = |\mathcal{K}|$ implies $\psi_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}$ is bijective. Thus, $\mathcal{S}$ is isomorphic to $\mathcal{K}$. Also observe that

$$\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}}(s_{i_1}, s_{i_2}, \ldots, s_{i_t}) = \prod_{i \in \mathcal{B}} \psi_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}(s_i) \, . \tag{4}$$

Let $\mathcal{K}$ be a finite Abelian group such that $\mathcal{K} \cong Z_2 \times Z_{q_2^{\omega_2}} \times \cdots \times Z_{q_c^{\omega_c}}(+)$ where $q_h \neq 2$ are primes and ω_h are integers > 0 $(2 \le h \le c)$. We denote the identity in this group as 0. We prove by contradiction that when $l > 2$ there is no ideal homomorphic threshold scheme over such $\mathcal{K}$. Assume that there is one, then to $s \in \mathcal{S}$ corresponds a (s', s'') where $s' \in Z_2$ and $s'' \in Z_{q_2}^{\omega_2} \times \cdots \times Z_{q_c}^{\omega_c}$, similarly $k \in \mathcal{K}$ corresponds to (k', k''). So $\eta_{\mathcal{B},\mathcal{X}_\mathcal{A}} : \mathcal{S}^t \to \mathcal{K}$ corresponds to $\eta'_{\mathcal{B},\mathcal{X}_\mathcal{A}} : (Z_2 \times (Z_{q_2^{\omega_2}} \times \cdots \times Z_{q_c^{\omega_c}}))^t \to Z_2 \times (Z_{q_2^{\omega_2}} \times \cdots \times Z_{q_c^{\omega_c}})$ and similarly we can define $\psi'_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}$. Because $\psi'_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}$ gives a group isomorphism we have $\psi'_{i,\mathcal{B},\mathcal{X}_\mathcal{A}}((s'_i, 0)) = (k'_i, 0)$ and using (4) we obtain that $\eta'_{\mathcal{B},\mathcal{X}_\mathcal{A}}((s'_{i_1}, 0), \ldots, (s'_{i_t}, 0)) = (k', 0)$. Similarly $\eta'_{\mathcal{B},\mathcal{X}_\mathcal{A}}((0, s''_{i_1}), \ldots, (0, s''_{i_t})) = (0, k'')$.

Now when $\eta'_{\mathcal{B},\mathcal{X}_\mathcal{A}}((s'_{i_1}, s''_{i_1}), \ldots, (s'_{i_t}, s''_{i_t})) = (a, b)$ then $a = k' \bmod 2$ and $b = k''$ in $Z_{q_2^{\omega_2}} \times \cdots \times Z_{q_c^{\omega_c}}$, because $\eta'_{\mathcal{B},\mathcal{X}_\mathcal{A}}$ is a function. So this induces a threshold scheme over $\mathcal{K}' \cong Z_2$ with $l > 2$. According to [19], the maximum l in any threshold scheme is $l_{\text{MAX}} \leq |\mathcal{K}| + t - 2$, so we have a contradiction. □

Theorem 7.2 shows that some homomorphic threshold schemes cannot be ideal. Perfect sharing schemes for some access structures have similar properties. The problem for general access structures was first addressed in [2]. Its is also known that for some t and l no ideal threshold schemes exist [19].

In the proof of Theorem 7.2 we used Z_2; using Z_p with $p \neq q_i$ gives a generalization. Theorem 7.2 leads to the question: What is the optimal size of the shares in homomorphic threshold schemes and in zero-knowledge homomorphic threshold schemes (which is $O(l) \cdot log_2|\mathcal{K}|$ in our zero-knowledge scheme). We now discuss whether the size of the shares in our scheme can be reduced.

Our zero-knowledge threshold scheme is related to the Lenstra constant [15], [23]. Let $Z[u]$ be an algebraic extension of the integers. The Lenstra constant $L(Z[u])$ is the cardinality of the largest set $\mathcal{U}$ of integers in $Z[u]$ whose nonzero differences are units. (The Lenstra constant was developed to provide a method to find new Euclidean fields, as used [15], [20], [21].) Observe that the set $\mathcal{X}_\mathcal{A} \cup \{0\}$, where $\mathcal{X}_\mathcal{A}$ is the set of x_i for the zero-knowledge threshold scheme, is the set $\mathcal{U}$ where $Z[u]$ is defined in §6. The scheme presented in §6 can be improved when for each positive integer l (l is as before, the total number of shareholders), an algebraic extension $Z[u]_l$ of small degree can be constructed such that $l < L(Z[u]_l)$. Finding such a family of algebraic extensions with small degree (in function of l) and large Lenstra constant is unfortunately an open problem [23]. Comparing the approach followed in §6 with the state of the art [21], the degree of the extension can only be reduced by a factor of two. So it seems to be difficult to reduce the size of shares of our scheme.

8. Conclusion.

We introduced zero-knowledge sharing schemes and motivated them. Existing perfect sharing scheme have been studied over finite fields or over some finite geometry. We have presented a homomorphic perfect zero-knowledge threshold scheme which works for any finite Abelian group provided that l is polynomial (to be more precise $l = O(|n|^c)$). Whether such schemes exist for any l is a natural open problem.

Although not necessarily zero-knowledge, we have proposed a homomorphic perfect threshold schemes over any finite Abelian group for any l. The distributor must know the exponent of the group for these schemes. When zero-knowledge is of little importance, e.g., when the exponent is public, these schemes have certain advantages such as requiring less memory. This introduces the following open problems. Can the efficiency of the schemes presented in this paper be improved? We proved that ideal homomorphic threshold schemes over some Abelian groups do not exist and we argued that reducing the size of shares seems a difficult problem if a similar approach as in this paper is followed. Finally, what other algebraic structures allow for homomorphic perfect (zero-knowledge) sharing schemes?

A. Appendix.

The following lemma is used in Corollary 6.2.

LEMMA A.1. *The elements* $h_m = \sum_{i=0}^{m-1} u^i$ *for* $1 \leq m \leq q-1$ *have an inverse in* $Z[u]/(u^{q-1} + \cdots + 1)$ *of the form* $\sum_{i=0}^{q-2} b_i u^i$ *where* $b_i \in \{-1, 0\}$ *or* $b_i \in \{0, 1\}$.

Proof. Observe that h_1 and h_{q-1} clearly satisfy the condition. For the rest of the proof we assume $2 \leq m \leq q-2$. To prove this lemma we first create the regular matrix representation [18, p. 424], $\rho(h_m)$. The inverse of h_m is represented by a vector

$\vec{h}'_m = (a_0, \ldots, a_{q-2})$ such that $\rho(h_m)\vec{h}'^T_m = (1, 0, \ldots, 0)^T$ over the rationals where T indicates the transpose. Let $p = q$ and $n = m$. It can be verified that $\rho(h_m)$ is $(b_{i,j})$ with $0 \le i \le p-2$ and $0 \le j \le p-2$ where

$$b_{i,j} = \begin{cases} 1 & \text{for } j = 0, \ldots, p-(n+1) \text{ and } i = j, \ldots, j+n-1, \\ -1 & \text{for } j = p-n, \ldots, p-2 \text{ and } i = j-p+n, \ldots, j-1, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

So we start to solve $(b_{i,j}) \cdot \vec{k}'^T = (1, 0, \ldots, 0)^{\mathrm{T}}$ where $\vec{k}' = \vec{h}'_m$. We now simplify (5) by reducing the number of unknowns and equations. Observe that $a_{p-(n+1)} = 0$ since $b_{p-2,i} = 0$ for $i \neq p-(n+1)$. Also $a_{p-2-i} = a_{p-(n+2)-i}$ for $0 \le i < n-1$ when $n < p-n$ since subtracting row $p-2-i$ from $p-3-i$ gives the equation $a_{p-2-i} - a_{p-(n+2)-i} = 0$ for $0 \le i \le n-2$. With these conditions we can make a new set of equations and have as unknowns $\vec{k}'_m = (a_0, \ldots, a_{p-(n+2)})$ and $p-(n+1)$ equations. The new matrix $(b_{i,j})$, corresponding to the new set of equations, satisfies (5) except p is now $p-n$, so $p := p-n$.

When $n > p-n$, the number of unknowns and equations may be reduced in another way. We note that $a_{p-(n+1)} = 0$ for the same reason as before. This time $a_{p-2-i} = a_{p-(n+2)-i}$ for $0 \le i < p-(n+1)$ for a similar reason as before. With these conditions we can make a new set of $n-1$ equations and have as unknowns $\vec{k}'_m = (-1 \cdot a_{p-n}, \ldots, -1 \cdot a_{p-2})$. By multiplying the so obtained matrix by -1 and rearranging the columns we can create a new matrix $(b_{i,j})$ satisfying (5) except p is now n and n is now $p-n$, so $tmp := p$, $p := n$, $n := tmp - n$.

These reductions on the matrix will be performed until n becomes 1 or $p-1$. It is obvious that the remaining a_i are 1 or 0. Observe that p and n are reduced in the same way as in the calculations of $\gcd(p, n)$ using the "primitive" Euclidean algorithm. This explains why the process of reduction terminates. □

Acknowledgments. The authors are grateful to Alfredo De Santis for discussions related to this paper, which were helpful to prepare this text. The authors wish to thank Hendrick Lenstra, Armin Leutbecher, and Gerhard Niklasch for discussions of the Lenstra constant. We appreciate that Josh Benaloh, Bob Blakley, Gus Simmons, and Moti Yung have expressed their interest in this work.

REFERENCES

[1] J. C. Benaloh, *Secret sharing homomorphisms: Keeping shares of a secret secret*, in Advances in Cryptology, Proc. Crypto '86 (Lecture Notes in Computer Science 263), A. Odlyzko, ed., Springer-Verlag, New York, Berlin, 1987, pp. 251–260.

[2] J. C. Benaloh and J. Leichter, *Generalized secret sharing and monotone functions*, in Advances in Cryptology, Proc. Crypto '88 (Lecture Notes in Computer Science 403), S. Goldwasser, ed., Springer-Verlag, New York, Berlin, 1990, pp. 27–35.

[3] E. R. Berlekamp, *Algebraic Coding Theory*, McGraw–Hill Book Company, New York, 1968.

[4] ———, *Factoring polynomials over large finite fields*, Math. Comp., 24 (1970), pp. 713–735.

[5] G. R. Blakley, *Safeguarding cryptographic keys*, in Proc. Nat. Computer Conf. AFIPS Conf. Proc., 1979, pp. 313–317.

[6] G. R. Blakley and L. Swanson, *Infinite structures in information theory*, in Advances in Cryptology. Proc. Crypto '82, D. Chaum, R. Rivest, and A. T. Sherman, eds., Plenum Press, New York, 1983, pp. 39–50.

[7] E. Brickell and D. R. Stinson, *The detection of cheaters in threshold schemes*, SIAM J. Discrete Math., 4 (1991), pp. 502–511.

[8] D. Coppersmith, A. Odlyzko, and R. Schroeppel, *Discrete logarithms in GF(p)*, Algorithmica, 1 (1986), pp. 1–15.

[9] Y. DESMEDT AND Y. FRANKEL, *Shared generation of authenticators and signatures*, in Advances in Cryptology Proc. Crypto '91, (Lecture Notes in Computer Science 576), J. Feigenbaum, ed., Springer-Verlag, New York, Berlin, 1992, pp. 457–469.

[10] Y. DESMEDT AND Y. FRANKEL, *Perfect zero-knowledge sharing schemes over any finite Abelian group*, in Sequences II (Methods in Communication, Security, and Computer Science), R. Capocelli, A. D. Santis, and U. Vaccaro, eds., Springer-Verlag, New York, Berlin, 1993, pp. 369–378.

[11] Z. GALIL, S. HABER, AND M. YUNG, *Minimum-knowledge interactive proofs for decision problems*, SIAM J. Comput., 18 (1989), pp. 711–739.

[12] R. G. GALLAGER, *Information Theory and Reliable Communications*, John Wiley and Sons, New York, 1968.

[13] J. GILL, *Computational complexity of probabilistic Turing machines*, SIAM J. Comput., 6 (1977), pp. 675–695.

[14] S. GOLDWASSER, S. MICALI, AND C. RACKOFF, *The knowledge complexity of interactive proof systems*, SIAM J. Comput., 18 (1989), pp. 186–208.

[15] J. H. W. LENSTRA, *Euclidean number fields of large degree*, Invent. Math., 38 (1977), pp. 237–254.

[16] G. HARDY AND E. WRIGHT, *An Introduction to the Theory of Numbers*, 5th ed., Oxford Science Publications, London, Great Britain, 1985.

[17] M. ITO, A. SAITO, AND T. NISHIZEKI, *Secret sharing schemes realizing general access structures*, in Proc. IEEE Global Telecommunications Conf., Globecom 87, IEEE Communications Soc. Press, 1987, pp. 99–102.

[18] N. JACOBSON, *Basic Algebra I*, W. H. Freeman and Company, New York, 1985.

[19] E. D. KARNIN, J. W. GREENE, AND M. HELLMAN, *On secret sharing systems*, IEEE Trans. Inform. Theory, 29 (1983), pp. 35–41.

[20] A. LEUTBECHER, *Euclidean fields having a large Lenstra constant*, Ann. Inst. Fourier Grenoble, 35 (1985), pp. 83–106.

[21] A. LUETBECHER AND G. NIKLASCH, *On cliques of exceptional units and Lenstra's construction of Euclidean fields*, in Number Theory, Ulm 1987 (Lecture Notes in Mathematics 1380), H. Schlickewei and E. Wirsing, eds., Springer-Verlag, New York, Berlin, 1988, pp. 150–178.

[22] A. MENEZES, S. VANSTONE, AND T. OKAMOTO, *Reducing elliptic curve logarithms to logarithms in a finite field*, in Proc. 23rd Annual ACM Symp. Theory of Computing, STOC, 1991, pp. 80–89.

[23] W. NARKIEWICZ, *Elementary and Analytic Theory of Algebraic Numbers*, 2nd ed., Springer-Verlag, New York, 1990.

[24] A. M. ODLYZKO, *Discrete logs in a finite field and their cryptographic significance*, in Advances in Cryptology, Proc. Eurocrypt 84 (Lecture Notes in Computer Science 209), N. C. T. Beth and I. Ingemarsson, eds., Springer-Verlag, New York, Berlin, 1984, pp. 224–314.

[25] M. RABIN, *Digitalized signatures and public-key functions as intractable as factorization*, Massachusetts Institute of Technology Technical Report MIT/LCS/TR–212, Cambridge, MA, January 1977.

[26] ———, *Probabilistic algorithms in finite fields*, SIAM J. Comput., 9 (1980), pp. 273–280.

[27] ———, *Efficient dispersal of information for security, load balancing, and fault tolerance*, J. Assoc. Comput. Mach., 36 (1989), pp. 335–348.

[28] R. L. RIVEST, A. SHAMIR, AND L. ADLEMAN, *A method for obtaining digital signatures and public key cryptosystems*, Comm. ACM, 21 (1978), pp. 294–299.

[29] A. SHAMIR, *How to share a secret*, Comm. ACM, 22 (1979), pp. 612–613.

[30] C. E. SHANNON, *Communication theory of secrecy systems*, Bell System Tech. Jour., 28 (1949), pp. 656–715.

[31] G. J. SIMMONS, *Robust shared secret schemes*, Congr. Numer., 68 (1989), pp. 215–248.

[32] G. J. SIMMONS, W. JACKSON, AND K. MARTIN, *The geometry of shared secret schemes*, Bulletin of the Institute of Combinatorics and its Applications, 1 (1991), pp. 71–88.

[33] D. R. STINSON AND S. A. VANSTONE, *A combinatorial approach to threshold schemes*, SIAM J. Discrete Math., 1 (1988), pp. 230–236.