SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

THE ALGEBRAIC GEOMETRY OF MOTIONS OF BAR-AND-BODY FRAMEWORKS*

NEIL WHITE† AND WALTER WHITELEY‡

Abstract. This paper generalizes and extends previous results on bar-and-joint frameworks to bar-and-body frameworks: structures formed by rigid bodies in space linked by rigid bars and universal joints. For a multi-graph which can form an isostatic (minimal infinitesimally rigid) bar-and-body framework, a single polynomial—the pure condition—is found which describes those bad positions of the bars for which infinitesimal rigidity fails. (The proof is much shorter than the previous derivation for bar-and-joint frameworks and the condition is linear in the variables.) The pure condition is used to describe the infinitesimal motions of a 1-underbraced framework in terms of the screw centers of motion of the bodies. The factoring of the polynomial condition is given by the lattice of isostatic blocks in the framework, with at most one irreducible factor for each block. For frameworks realized at generic points of an irreducible factor the infinitesimal motions and the static stresses are also given by the factoring and the lattice. (These results are much sharper than the corresponding results for bar-and-joint frameworks.) The theorems are presented in terms of k-frames—a simple generalization of bar-and-body frameworks which also has applications to scene analysis and other types of frameworks.

Key words. bar-and-body frameworks, infinitesimal motions, static stress, polynomial conditions, irreducible factor, lattice of blocks

AMS(MOS) subject classifications. Primary 51M30, 70B99, 70C99; secondary 73K99, 68G10

Introduction. The traditional bar-and-joint frameworks have been generalized to bar-and-body frameworks, in which large rigid bodies are tied together with rigid bars, each attached to a pair of distinct bodies by universal joints. Several factors make these structures important:

(i) a number of critical problems in 3-space are unsolved for bar-and-joint frameworks in 3-space;

(ii) the analogous problems are solved in all dimensions for bar-and-body frameworks;

(iii) the results for bar-and-body frameworks apply directly to hinged panel structures which are commonly built in 3-space.

In particular, the problem of characterizing the graphs of isostatic (minimal infinitesimally rigid) bar-and-joint frameworks is unsolved for $n=3$ [23], while the characterization of the multi-graphs for isostatic bar-and-body frameworks has recently been solved for all dimensions (see § 2.4 and [21], [23], [29]).

In our first paper [25] we described the pure condition for the graph of an isostatic bar-and-joint framework in n-space—a single polynomial in the coordinates of the joints whose zeros identify the special realizations of the graph which are not infinitesimally rigid. These conditions were used to investigate the static behavior of overbraced frameworks, and preliminary work was done on the factoring of the conditions.

* Received by the editors January 2, 1985; accepted for publication (in revised form) February 21, 1986. This work was supported in part by grants from the Natural Sciences and Research Council of Canada and Fonds pour la formation de chercheurs et l'aide à la recherche, Quebec.

† Department of Mathematics, University of Florida, Gainesville, Florida 32611.

‡ Department of Mathematics, Champlain Regional College, St. Lambert, Quebec J4P 3P2, Canada.

Here we present the much simpler pure condition for the multi-graph of an isostatic bar-and-body framework—a single polynomial which is now linear in the variables for each bar. These conditions are derived in § 2, along with a technique for their computation based on work of Rosenberg [16] and an application to the characterization of the graphs of isostatic frameworks. In § 3 the pure conditions are used to describe the screw centers for the infinitesimal motions of the bodies in a 1-underbraced bar-and-body framework. This work uses the projective algebra of screws which dates back to the last century [11] but has also been revived for work in invariant theory [1], [5] and in robotics [9], [13], [20].

For the multi-graph of an isostatic framework the isostatic subpieces (or blocks) form a lattice and these blocks correspond precisely to the factoring of the pure condition of the graph: each irreducible factor is associated with a unique such block, and no block has more than one factor (Theorem 4.12). Each edge is associated with the lowest block in which it occurs, and this partition describes which edges lie in the static stress at any generic point of an irreducible factor (namely those edges in blocks at or below the block of the factor (Proposition 4.6)) and which edges join bodies in motion relative to each other at a generic point of the factor (those edges associated with blocks at or above the block of the factor (Proposition 4.8)). We conjecture that this same pattern of irreducible factors, edges and blocks applies to bar-and-joint frameworks in the plane, although any proofs will have to be more complex. We do know that the pattern must be modified for bar-and-joint frameworks in 3-space.

Bar-and-body frameworks are really special examples of a general matroid structure we call a k-frame [29]. Accordingly we present the results of the paper in terms of these general structures. The k-frames first appeared in scene analysis—the study of hyperplane scenes in R^k projected into pictures in R^{k-1} [28]. In § 5 we briefly describe this interpretation for 3-frames to indicate how our results apply in this field. In passing, we note that the k-frames can also be used to describe bar-and-joint frameworks on the torus T^k (a quotient of R^k by the unit hypercube) [29]. As a consequence, k-frames, and our work here, have potential applications to the study of periodic sphere packings in R^k [3].

A number of the techniques we use are based on the form of the rigidity matrix for the framework or k-frame. This form, in turn, reflects the underlying matroid structure of the matroid union of k copies of the graphic matroid [29]. As a result, these techniques also apply to many other represented matroids based on unions of graphic and bicircular matroids of a graph [29]. More generally, we anticipate that many of the techniques will apply to a large class of matroids defined by counts on a graph or hypergraph [26], [30].

1. Introduction to bar-and-body frameworks and frames. Let B be a rigid body in R^n. Then any instantaneous motion of B may be expressed as a vector sum of rotations and translations of B, as is well known. For example, in R^2, any such motion is a rotation or a translation, and in R^3, such a motion is in general a screw, or a rotation about a line L plus a translation in a direction parallel to L. In R^4, there are motions which cannot be expressed more simply than as a sum of two rotations. We must first develop the algebra of such motions. We give an informal presentation of this algebra; more details may be found in [4], [27].

1.1. Centers of motions in n-space. To any point p in R^n we will assign the homogeneous coordinates $(p_1, p_2, \cdots, p_n, 1)$. Thus we are regarding R^n as embedded in projective space $PG(R, n)$. Since rigidity properties are in general projectively invariant [17], it is useful and sometimes necessary to work with arbitrary subspaces in $PG(R, n)$. Such a subspace W of dimension d corresponds to a subspace of dimension $d+1$ in R^{n+1}. We say that W has rank $d+1$, and recall that it takes $d+1$ points to determine such a subspace. Thus, for example, an arbitrary line in R^n (plus its point at infinity) is a subspace of rank 2, but so is any line at infinity in $PG(R, n)$.

Now let the subspace W of rank r be determined by the points $p^1, p^2, \cdots, p^r$ (i.e., $p^2, p^2, \cdots, p^r$ is actually a *basis* of W in R^{n+1}). In the Cayley algebra on R^{n+1} (see [5], also referred to as Peano spaces on R^{n+1} in [1]), we may associate with W the *r-extensor* $p^1 \vee p^2 \vee \cdots \vee p^r$. This r-extensor is, in the coordinatized version, really just the vector of Plücker coordinates of W, that is, the sequence of $r \times r$ minors of the $r \times (n+1)$ matrix whose rows are $p^1, p^2, \cdots, p^r$.

Now we consider a rotation of B in R^n, or what is really equivalent, a rotation of all of R^n itself. Any such rotation has a center (or axis) which is a subspace W of rank $n-1$. Let $Z'' = p^1 \vee p^2 \vee \cdots \vee p^{n-1}$ be the associated $(n-1)$ extensor. Then for any point p not in W, $Z'' \vee p$ is an n-extensor associated with the hyperplane sp $(W+p)$. Furthermore, $Z'' \vee p$ is an $(n+1)$-vector whose entries are the coefficients of the equation of the hyperplane sp $(W+p)$ (assuming certain sign and order conventions). That is, the first n coordinates are the vector v normal to sp $(W+p)$ (the $(n+1)$st entry being the constant term $-v \cdot (p_1, \cdots, p_n)$). If p' is another point in sp $(W+p)$, then $Z'' \vee p$ and $Z'' \vee p'$ are scalar multiples of each other in the same ratio as the ratio of the distances of p and p' from W (with opposite signs if they are on opposite sides of W in sp $(W+p)$). Thus, for some constant scalar α, $\alpha(Z'' \vee p)$ is (except for its last entry) the velocity vector of the rotation at p, for every point p. We may regard α as the angular velocity, appropriately normalized.

We will henceforth refer to the $(n-1)$-extensor $Z' = \alpha Z''$ as the *center* of the rotation, and for any point p, $M(p) = Z' \vee p$ as the *motion* at p.

Next we consider a translation in the fixed direction of the free vector v. Let U be any hyperplane in the parallel family of hyperplanes normal to v. Then U intersects the hyperplane at infinity (=points with $(n+1)$st coordinate zero in $PG(R, n)$) in a subspace W of rank $n-1$, where W is independent of the choice of U. Regarding W as the center of a "rotation," we mimic the above development. If Z'' is the $(n-1)$-extensor corresponding to W and α the chosen scalar for our translation, then $Z' = \alpha Z''$ is the *center* and $M(p) = Z' \vee p$ is the *motion* at p. This also corresponds to the equation of the hyperplane normal to the velocity, and for an appropriate scalar α it is the velocity vector v (independent of p) together with one extra component, $-v \cdot (p_1, \cdots, p_n)$.

If we now take an arbitrary instantaneous motion of our rigid body, it is a vector sum of rotations and translations. If $Z'_1, Z'_2, \cdots, Z'_m$ are the centers of these rotations and translations, then $Z = \Sigma Z'_i$ is the *center* of our motion and $M(p) = \Sigma Z'_i \vee p = Z \vee p$ is the *motion* at p for any point p of the rigid body. $M(p)$ still represents the coordinates of the hyperplane through p normal to the velocity vector.

Remark 1.1. The linear combination Z of $(n-1)$-extensors is no longer an $(n-1)$-extensor, unless the motion is again a rotation or translation. The center Z may now be an arbitrary vector of length $\binom{n+1}{n-1} = \binom{n+1}{2}$, whereas an $(n-1)$-extensor satisfies the Grassmannian quadratic relations (see [8, pp. 309-315]). The study of screw motions in R^3 from this point of view has an extensive literature and is of current interest in the study of robotics [9], [13], [20].

Example 1.2. Consider a rotation about the x-axis in R^3. Taking the origin $(0, 0, 0, 1)$ and the point $(1, 0, 0, 1)$ on the x-axis, we get as the center the Plücker coordinate vector $Z = (0, 0, -1, 0, 0, 0)$, here taken in the order $P_{12}, P_{13}, P_{14}, P_{23}, P_{24}, P_{34}$, where P_{ij} denotes the minor using columns i and j. The motion vector $Z \vee p$, taken in the order $P_{234}, -P_{134}, P_{124}, -P_{123}$ for minors of

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ p_1 & p_2 & p_3 & 1 \end{bmatrix}$$

is $M(p)=(0, -p_3, p_2, 0)$, where $p=(p_1, p_2, p_3, 1)$. We note that $(0, -p_3, p_2)$ is correct for the velocity vector of our rotation, up to a scalar. Clearly the fourth coordinate is 0 because the plane $M(p)$ contains the origin.

Similarly, for a translation in the direction of the x-axis, $v=(1,0,0)$, U may be taken as the hyperplane defined by $x=0$, and W is the line at infinity $\{(x, y, z, w)|x=w=0\}$. This line is determined by the points at infinity $(0,1,0,0)$ and $(0,0,1,0)$, hence $Z=(0,0,0,1,0,0)$, and $M(p)=(1,0,0,-p_1)$.

1.2. Bar-and-body frameworks. Suppose now that two rigid bodies, B_1 and B_2, are connected by a rigid bar, which is attached flexibly at points a and b on B_1 and B_2 (resp.). If B_1 and B_2 undergo instantaneous motions with centers Z_1 and Z_2 (resp.), then the condition that the distance from a to b is instantaneously preserved is the following:

If u and v are the velocity vectors at a and b,

$$M(a)=(u, -u\cdot(a_1, \cdots, a_n)), \qquad M(b)=(v, -v\cdot(b_1, \cdots, b_n));$$

then

$$\begin{aligned} 0 &= (u-v)\cdot((a_1, \cdots, a_n)-(b_1, \cdots, b_n)) \\ &= u\cdot(a_1, \cdots, a_n)-u\cdot(b_1, \cdots, b_n)-v\cdot(a_1, \cdots, a_n)+v\cdot(b_1, \cdots, b_n) \\ &= -M(a)\vee b-M(b)\vee a \end{aligned}$$

or

$$Z_1 \vee a \vee b+Z_2 \vee b \vee a=Z_1 \vee (a \vee b)-Z_2 \vee (a \vee b)=(Z_1-Z_2)\vee(a \vee b)=0$$

(see [4]). Here $a \vee b$ is a 2-extensor, a vector of length $\binom{n+1}{2}$ consisting of the 2×2 minors (Plücker coordinates) of the $2\times(n+1)$ matrix whose rows are a and b. We will henceforth write $a \vee b$ simply as ab.

If Z is an $(n-1)$-extensor $p^1 \vee p^2 \vee \cdots \vee p^{n-1}$, let us take $Z^*=(P'_{12}, -P'_{13}, \cdots, (-1)^{i+j-1}P'_{ij}, \cdots)$, where P'_{ij} is the $(n-1)\times(n-1)$ minor obtained by omitting columns i and j from the matrix whose rows are $p^1, p^2, \cdots, p^{n-1}$. Now if the 2-extensor ab is written in the standard order $(Q_{12}, Q_{13}, \cdots, Q_{ij}, \cdots)$, where Q_{ij_1} is the minor with columns ij, then

$$\begin{aligned} Z \vee ab &= p^1 \vee p^2 \vee \cdots \vee p^{n-1} \vee a \vee b=\det(p^1, p^2, \cdots, p^{n-1}, a, b) \\ &= \pm(\Sigma\,(-1)^{i+j-1}P'_{ij}Q_{ij})=\pm Z^*\cdot ab \end{aligned}$$

where a Laplace expansion was used on the last two rows of the determinant. We have proved:

PROPOSITION 1.3. *If rigid bodies B_1 and B_2 undergo motions with centers Z_1 and Z_2 (resp), then the length of a bar ab is preserved if and only if*

$$Z_1^*\cdot ab-Z_2^*\cdot ab=0.$$

DEFINITION 1.4. A *bar-and-body framework* in R^n is a finite collection of disjoint rigid bodies $B_1, B_2, \cdots, B_m$ and of rigid bars $(a_1, b_1), (a_2, b_2), \cdots, (a_e, b_e)$, where a_i and b_i are points on distinct bodies and the ith bar is attached flexibly to those two points as its end points.

We will assume that each body is *full*, that is, it spans an affine subspace of R^n of at least rank n (i.e. dimension $n-1$). Corresponding to any such framework is a finite graph G with vertices corresponding to the bodies and edges to the bars. We may have multiple edges in G but no loops. Any such graph G may be realized as a framework by assigning an ordered pair of $(n+1)$-tuples (with last entry equal to one) to each edge, and providing that points on distinct bodies are assigned distinct $(n+1)$-tuples. We note that the size and shape of the bodies themselves are not relevant to questions of instantaneous motions, provided each body contains the requisite end points of bars. We also note that the various edges incident to the ith vertex of G may be given distinct end points on the ith body B_i. We denote by $G(p)$ any such particular realization of the graph G as a bar-and-body framework.

DEFINITION 1.5. The *rigidity matrix* $M(G(p))$ for the framework $G(p)$ has one row for each bar and $\binom{n+1}{2}$ columns for each body, with the columns for B_1 followed by those for B_2, etc. If (a, b) is a bar with end points a in body B_i and b in body B_j, then the row corresponding to (a, b) in $M(G(p))$ has the 2-extensor ab in the $\binom{n+1}{2}$ columns for B_i, $-ab$ in the $\binom{n+1}{2}$ columns for B_j, and 0 in all other columns. (Under this definition, many frameworks are equivalent. Indeed, all that matters is the 2-extensor, or line, ab, not the location of the two points a and b on that line.)

A *motion* of $G(p)$ is an assignment of a center Z_i to each body B_i, $1 \leqq i \leqq m$, so that the length of each bar (a, b) is instantaneously preserved, that is, $Z_i^* \cdot ab - Z_j^* \cdot ab = 0$. If we let Z^* be the vector of length $m\binom{n+1}{2}$ consisting of Z_1^* followed by Z_2^* followed by Z_3^*, etc., then we require that $Z^* \cdot R = 0$ for each row R of $M(G,(p))$.

PROPOSITION 1.6. *Motions of the framework $G(p)$ correspond (under *) to the orthogonal subspace to the row-space of $M(G(p))$, i.e., to solutions of $M(G(p)) \times Z^{*t} = 0$, where ^{t} denotes transpose.*

Now the Euclidean motions of all of R^n, obtained by setting all Z_i equal to each other, are always motions of $G(p)$. Since these motions form a subspace of dimension $k = \binom{n+1}{2}$, the maximum rank of $M(G(p))$ is $(m-1)k$. We say $G(p)$ is *isostatic* (or *basic*) if $M(G(p))$ has exactly $(m-1)k$ rows which are linearly independent.

1.3. k-frames.

For the remainder of this paper, we wish to adopt a more general point of view, by working with k-frames rather than bar-and-body frameworks. The concept of k-frame includes bar-and-body frameworks as a special case, but also includes applications to scene analysis (see § 5) and other types of frameworks [29], [30].

DEFINITION 1.7. Let G be a graph with no loops but possibly with multiple edges. A *k-frame matrix* for G consists of one row for each edge and k columns for each vertex, where if $e = (u, v)$ is an edge of G, then the row for e has a k-tuple x_e in the columns for u, $-x_e$ in the columns for v, and 0 in all other columns. This matrix, for any particular choice of a vector x_e for each edge e, is denoted $M(G(p))$, and the graph G together with such assignments of x_e is called a *k-frame* $G(p)$. If $G(p)$ has distinct algebraically independent indeterminates for all entries in the x_e's, we call $G(p)$ a *generic k-frame* for G, and denote the corresponding k-frame matrix $M(G)$.

A *motion* of a k-frame $G(p)$ is a vector Z^* of length km which is orthogonal to the row space of $M(G(p))$. The *trivial motions*, having the same k-tuple for each vertex, are always motions of $G(p)$. A k-frame is *rigid* if it has only the trivial motions.

We note that bar-and-body frameworks are the special case of k-frames in which $k = \binom{n+1}{2}$ and the vectors x_e are all 2-extensors. It is possible to interpret more general k-frames as situations similar to bar-and-body frameworks. For example, 2-frames may be interpreted as bar-and-body frameworks on a cylinder or torus, where two independent directions of "translation" are allowed, one of which is rotation about the axis of the cylinder, but no rotation of a body about a point of the body is allowed [29]. We will not pursue this interpretation further.

2. The pure condition for a graph.

2.1. Tie-downs and the pure condition. We want an algebraic procedure to describe which positions (if any) of the edges of the graph as a k-frame will give an independent, or spanning set. The graphs which give the simplest formulae are those which at least count to be a maximal independent set as a k-frame. Since any k-frame has a k-dimensional space of trivial motions we give the following definition.

DEFINITION 2.1. A graph is k-counted if $|E| = k|V| - k$.

Not all k-counted graphs will give the desired independent sets, but this is a necessary condition for the algebra. In Theorem 2.18 we describe the necessary and sufficient conditions on a graph.

It is a simple matter to check the independence, and the span, of the rows of a square matrix by taking the determinant. However the k-frame matrix for a generic k-frame on a k-counted graph is not square. We must add k simple rows which will square up the matrix and be independent of the rows for any k-frame.

DEFINITION 2.2. The *basic tie-down* Tm is a set of k rows and km columns of the form:

$$Tm = [I_k 0 \cdots 0].$$

For a framework in n space, $(k = n(n+1)/2)$, we can think of these rows as bars from the first body to the ground, designed to remove the trivial or Euclidean motions of the entire framework.

LEMMA 2.3. *The rows of Tm are independent of the rows of $M(G(p))$ for any k-frame on a graph with m vertices.*

Proof. Any k-frame matrix $M(G(p))$ has the k-dimensional space of trivial solutions generated by e_i^m—the vector formed by repeating the basic vector $e_i = (0, \cdots, 0, 1, \cdots, 0)$ m times. We claim that the k rows of Tm remove this k-space of solutions. In particular, row i removes e_i^m from the solution space.

Since each row removes a new solution, these rows are independent of the rows of $M(G(p))$ (and one another). □

For any k-frame $G(p)$ the matrix formed by adding the appropriate tie-down rows T (with k rows and $k|V|$ columns) to the bottom of the k-frame matrix is written $M(G(d), T)$.

DEFINITION 2.4. A k-frame $G(p)$ is *k-isostatic* if every allowed k-motion is trivial and deleting any edge introduces a nontrivial k-motion.

PROPOSITION 2.5. *If G is k-counted then any k-frame $G(p)$ is k-isostatic if and only if* $\det(M(G(p), T)) \neq 0$.

Proof. If $\det(M(G(p), T)) \neq 0$ then the rows of the matrix are of rank $k|V|$, and the rows of $M(G(p))$ are of rank $k(|V| - 1)$. Since the trivial k-motions form a space of dimension k, we conclude that the space of allowed motions is the space of trivial motions.

If $G(p)$ is k-isostatic then the rows of $M(G(p))$ are of rank $k(|V| - 1)$ (since the nullity is k). When we add T, we obtain a square matrix with rank $k|V|$ (by Lemma 2.3), so we conclude that $\det(M(G(p), T)) \neq 0$. If any edge is deleted, we have less than $k|V| - 1$ rows and there must be a nontrivial motion. □

We can describe the algebraic form of $\det(M(G, T))$.

PROPOSITION 2.6. *For any k-counted graph G*

$$\det(M(G, T)) = C(G)$$

where

(i) *$C(G)$ is a polynomial in the algebra of k-brackets of edges of G.*

(ii) *$C(G)$ is homogeneous of degree $|V|-1$ in the brackets.*

(iii) *$C(G)$ is linear in the variables of each edge vector.*

Proof. We start with a Laplace expansion on the first k columns of the matrix. Because the last k rows for T are zero outside these columns, we have a single term

$$\det(M(G,T)) = \det[I_k]\circ C(G) = C(G).$$

We now expand this minor $C(G)$ by a Laplace expansion on the k columns of the second vertex. This gives a sum of terms

$$\pm[b_{i1}\cdots b_{ik}]\circ C_{i1}\cdots{}_{ik}(G)$$

where the first factor is a k-bracket (or $k\times k$ determinant) with rows for k edges of the graph.

We repeat this decomposition, working through all columns k at a time, to obtain the required polynomial in the brackets. Each term is degree $|V|-1$ in the brackets.

Since each row can only be used once in each term of such a Laplace expansion, each term has exactly one entry $\pm b_i$ for each edge vector b_i—which is the desired linearity. □

COROLLARY 2.7. *For any k-frame $G(p)$ of a k-counted graph*:

$$C(G(p))\neq 0 \quad \textit{if and only if } G(p) \textit{ is k-isostatic.}$$

It appears that the polynomial $C(G)$—the *pure condition* of the graph in k-space—depends on the choice of k rows for T. Surprisingly any similar set of k rows would give the same critical factor $C(G)$.

PROPOSITION 2.8. *For any k-isostatic graph G with r vertices, and any set S of k rows of length k.r*

$$\det(M(G,S)) = f(S)\circ C(G)$$

where $f(S)$ is a polynomial in the entries of S and $f(S)\neq 0$ if the rows of S block the trivial motions.

Proof. Consider any assignment P of complex numbers p to the variables for edges in $M(G)$. The rows of $M(G(p))$ are dependent if $C(G(p))=0$. Therefore,

$$C(G(p))=0\rightarrow \det(M(G(p),S))=0.$$

Since we have two polynomials, and the implication holds for all complex numbers, Hilbert's Nullstellensatz [7, p. 165] guarantees that

$$(\det(M(G,S)))^n = A'\circ C(G).$$

However, $C(G)$ is of first degree in all variables so

$$C(C)|Q^n \text{ implies } C(G)|Q \quad\text{or}\quad \det(M(G,S)) = A\circ C(G).$$

Since $\det(M(G,S))$ has only one entry for each variable from G, we know that A is a polynomial only containing variables in S. We define $f(S)=A$.

If the rows of S block all trivial motions, then as in Proposition 2.5, $M(G,S)$ is invertible. Therefore

$$0\neq\det(M(G,S)) = f(S)\circ C(G) \quad\text{and}\quad f(S)\neq 0.$$

If the rows of S do not block all trivial motions, then $M(G,S)\times X=0$ has a nontrivial solution. Therefore $0=\det(M(G,S))=f(S)\circ C(G)$. Since $C(G)\neq 0$, we find $f(S)=0$. □

Remark 2.9. For a natural class of sets S we can describe $f(S)$. A generic tie-down of G is a set of k rows $r_1, \cdots, r_k$ such that for all but one vertex $V_{k(i)}$ all k entries of r_i are zero and the $V_{k(i)}$ entries are $Z_i = (z_{i1}, \cdots, z_{ik})$. It is not difficult to check that $f(S) = \pm[Z_1, \cdots, Z_k]$ for any generic tie-down.

If our k-frame is a bar-and-body framework, then each $(z_{i1}, \cdots, z_{ik})$ is a 2-extensor of a line (or bar) and $f(S)$ is a determinant of 2-extensors. This determinant is examined in more detail in [24].

Remark 2.10. The proof of the analogous theorem for bar-and-joint frameworks was far more complex [24, Prop. 3.12]. The simplicity of the current proof illustrates the advantages of the present type of structure.

2.2. A combinatorial formula for the pure condition. If we reexamine the basic Laplace decomposition which generated $C(G)$, we can give a precise graph theoretic description of which partitions of the edges in G give terms in the polynomial, as well as the signs of the terms. An analogous (but, naturally, more complex) description for bar-and-joint frameworks was given by Ivo Rosenberg [16]. For convenience we assume that all edges of the graph have been oriented in some arbitrary fashion.

DEFINITION 2.11. A *k-fan* for the graph G is a partition of the edges into disjoint ordered sets f_i, $2 \leqq i \leqq v$, such that each f_i is an ordered set of k edges all adjacent to the vertex v_i.

Two k-fans π and π' are *distinct* if f_i includes an edge not in f_i' for some i. Otherwise the two k-fans are *permutation equivalent.*

The *sign of a k-fan* π, written $\sigma(\pi)$, is the sign of the permutation from the ordered set E to the order $(f_2, f_3, \cdots, f_v)$ times $(-1)^r$ where r is the number of directed edges in E which are in the f_i of their second vertex.

As a matter of shorthand we write $[f_i]$ for the bracket $[c_1, \cdots, c_n]$ where c_j is the k-vector assigned to the jth edge in f_i.

PROPOSITION 2.12. *The pure condition of a k-counted graph G is*

$$C(G) = \Sigma\sigma(\pi)[f_2] \cdots [f_n] \qquad (\textit{sum over all distinct k-fans } \pi \textit{ of } G).$$

Proof. It is a simple matter to see that each nonzero term of the Laplace decomposition corresponds to such a k-fan. The actual bracket $[f_i]$ is, up to permutation, precisely the piece of the Laplace expansion term corresponding to the columns of vertex v_i. For those brackets, and the sign of the term in the Laplace expansion, we have the precise discrepancy $(-1)^r$ (to account for occurrences of $-b_j$, in the matrix when the edge enters v_j) and the permutation sign for the usual Laplace expansion rule. □

There is a simple and suggestive way to visually record a k-fan (or rather a permutation equivalence class of k-fans).

DEFINITION 2.13. The *k-fan diagram,* $D(\pi)$, is a directed graph with the vertices of G, but with all edges reoriented so that e_i is directed out of v_j if e_i is in f_j.

In Fig. 2.1A, B, C we illustrate the distinct 3-fans of a sample graph. As shown in Fig. 2.1D, we can move from k-fan diagram A to diagram B or C by reversing all edges of a directed polygon in the diagram. (A *directed polygon* is a cycle of edges and vertices such that all edges are directed around the cycle in the directed graph.)

Given a general k-fan π and a set of edge-disjoint simple directed polygons in the k-fan diagram $D(\pi)$, each polygon is reversed by replacing each edge (v_{ki}, v_{ki+1}) in a polygon, which was in f_i, by the edge (v_{ki}, v_{ki-1}). This creates a new set f_i' for each vertex, and creates a new k-fan π' called the *polygon reversal* of π.

PROPOSITION 2.14. *Given any two distinct k-fans π and π', there is a set of edge-disjoint, simple, directed polygons in the k-fan diagram $D(\pi)$ such that the polygon reversal of π on this set is a k-fan π'' which is permutation equivalent to π'.*

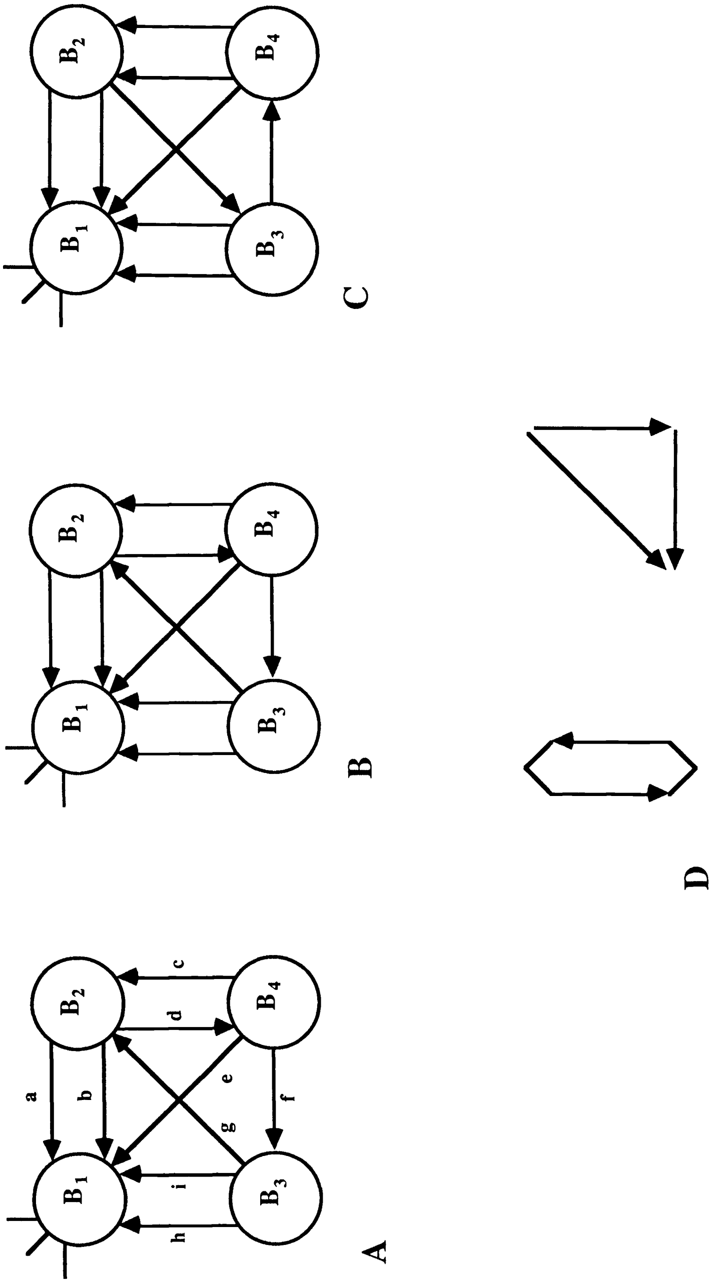


FIG. 2.1

Proof. Since the two k-fans are distinct, there is an f_i using an edge not in f'_i. This is the first edge in the path. This edge must be in f'_j (for its other end), and since f_j and f'_j are the same size, there is an edge in f_j not in f'_j. This is the second edge of the path. We repeat this process until the path repeats on a vertex. The loop between repetitions forms our first polygon.

We reverse this polygon, creating a new k-fan π'', which is closer to π' in the sense that $D(\pi'')$ and $D(\pi')$ have more edges with the same orientation.

If π'' and π' are still distinct, we repeat the process, creating additional polygons until we stop at a π'' such that $D(\pi'')=D(\pi')$. This means that π'' is permutation equivalent to π' as required.

Since no edge will be reversed twice, the process will terminate and the polygons are edge-disjoint. □

PROPOSITION 2.15. *Given a k-fan π and a polygon reversal π' obtained by reversing on r simple, edge-disjoint, directed polygons then*

$$\sigma(\pi')=(-1)^r\sigma(\pi).$$

Proof. Assume we reverse on one polygon of length m. The basic permutation for π' can be obtained from that for π by cycling the m edges—causing a sign change $(-1)^{m-1}$. However we have also switched these m bars in the count of bars entering their heads or tails in their f'_i—giving an additional sign change of $(-1)^m$. The total change is $(-1)^{m-1+m}=(-1)$.

If there are r polygons, the process is repeated r times and $\sigma(\pi')=(-1)^r\sigma(\pi)$. □

2.3. Examples of pure conditions. We will illustrate the techniques of the previous section by deriving the pure conditions for a few small examples.

Example 2.16. The graph illustrated in Fig. 2.2 has a unique 6-fan shown in Fig. 2.2B. As a result the pure condition is a single term. Assuming the edges of G are

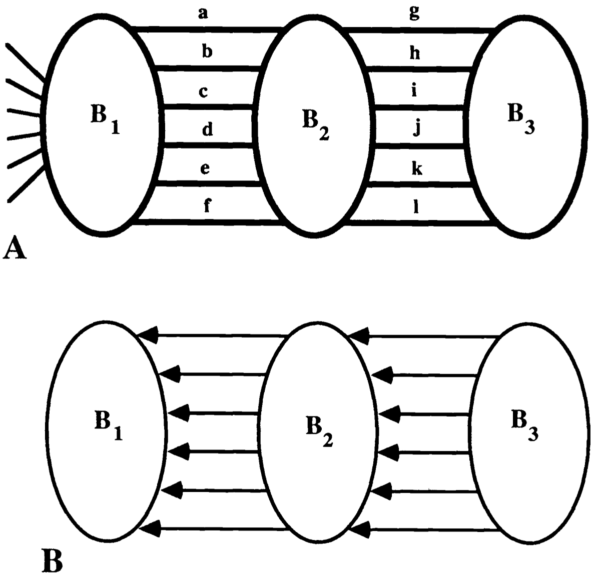


FIG. 2.2

oriented as in Fig. 2.2B and ordered lexicographically,

$$C(G)=[abcdef][ghijkl].$$

A graph with a unique k-fan covering is called *k-simple.*

For such simple examples $C(G)$ can also be easily obtained by a block decomposition of det $(M(G, T))$.

Example 2.17. We return to the graph illustrated in Fig. 2.1. Assuming the graph is oriented as in Fig. 2.1A, the three fans give the pure condition

$$C[G]=[abc][def][ghi]-[abd][cef][ghi]-[abg][dec][fhi].$$

2.4. A characterization of K-isostatic graphs. There are simple criteria for which graphs will have nonzero pure conditions as k-frames. We offer an alternate proof which illustrates the use of pure conditions, k-fans and a technique of specializing the lines of the bars.

THEOREM 2.18. *A graph G which is k-counted has a nonzero pure k-condition if and only if there is a set of k edge-disjoint spanning trees which cover the graph G, if and only if the rigidity matrix is the matroid union of k cycle matroids of the graph G.*

Proof. Assume the graph has a nonzero pure k-condition. With the standard tie-down we know det $(M(G(p), T))\neq 0$. We now reorder the columns of this matrix placing the $|V|=v$ columns of first entries for each vertex first, then the columns of second entries, etc. A Laplace decomposition following these blocks of v columns gives terms Π (det $N_i)1\leqq i\leqq k$ where each N_i is a square $v\times v$ matrix formed from the ith coordinates of the rows of v edges (possibly including "edges" in T).

There is at least one nonzero product Π det (N_i). By the Laplace construction, the rows used in the N_i form disjoint subgraphs, and it is clear, for a nonzero term, that each N_i will include the ith row from T. The remaining $v-1$ rows of N_i will each be of the form

$$\alpha_i[0\cdots 0\,1\,0\cdots 0\,-1\,0\cdots 0]$$

for some nonzero scalar α_i. Ignoring these factors, we have the usual matrix for the subgraph of these edges. Any polygon in this subgraph gives a simple linear dependence of the corresponding rows, so a nonzero term represents a subgraph with no polygons. Since we have v vertices, $v-1$ edges and no polygons, the subgraph must be a spanning tree T_i. Thus the distinct factors N_i give the required edge-disjoint spanning trees covering the edges of G.

Conversely, assume that G is covered by k edge-disjoint spanning trees. If we root all these trees at the first vertex, and direct all edges down towards this root, we have a k-fan diagram. Each vertex has k branches down to the root (one from each tree) and ordering these edges in the order of the trees gives a k-fan Π_0.

The pure condition is expressed

$$C(G)=\Sigma\sigma(\pi)[f_2][f_3]\cdots[f_v].$$

We specialize the vectors for the edges by assigning all edges from tree T_i the same set of indeterminates $\bar{X}_i=(x_{i1},\cdots,x_{ik})$. With this specialization $G(X)$, we have the term for π_0 (up to sign)

$$[\bar{X}_1\cdots\bar{X}_k]^{v-1}$$

since each f_i contains one edge from each tree.

Consider any k-fan diagram $D(\pi')$ which has only one out-directed edge from each tree. For each vertex there is a unique path to the root vertex in T_i. Since all edges to this first vertex are directed into it, and no vertex has two out-directed edges from T_i, this entire path is directed down to the root. This direction of edges matches the diagram for π, so we conclude that π' is permutation equivalent to π.

As a result any distinct k-fan has an f_i containing two edges from some tree. With the given specialization $[f_i]$ is zero because of the duplicate columns.

We can conclude that

$$C(G(X)) = \pm[\bar{X}_1 \cdots \bar{X}_n]^{v-1} \neq 0.$$

Since this specialization is $\neq 0$, the original polynomial is also nonzero. □

COROLLARY 2.19 (Tay [21]). *A graph is k-isostatic if and only if* $|E| = k(|V|-1)$ *and, for any nonempty subgraph* G': $|E'| \leqq k(|V'|-1)$.

Proof. By a theorem of Tutte and of Nash Williams a graph can be covered by k-edge-disjoint spanning trees if and only if it has the counts given [29]. □

Remark 2.20. Corollary 2.19 was first derived, for the case $k = n(n+1)/2$, by a very different proof. An alternate proof of Theorem 2.18, along with the extension of Tay to Grassmann coordinates of lines when $k = n(n+1)/2$ (describing actual bar and body frameworks in n-space) is also given in [29].

Remark 2.21. One value of the tree covering property of Theorem 2.18 lies in an efficient polynomial algorithm to find the trees. A direct verification of the counting property of Corollary 2.19 would use an exponential algorithm. The role of tree coverings in checking generic rigidity of bar and joint frameworks in the plane has been explored by Lovasz and Yemini [14] and by Recski [15].

Remark 2.22. The count of Corollary 2.19 can be indirectly checked by a second even more efficient algorithm introduced by Sugihara for plane bar-and-joint frameworks [18]:

> A graph satisfies $|E| = k|V| - k$ and $|E'| \leqq k|V'| - k$ for all nonempty subgraphs if and only if for each vertex i, when k tie-down edges are added at the vertex (as loops) to create G_i, $|E_i| = k|V_i|$ and, $|E_i'| \leqq k|V_i'|$ for all nonempty subgraphs.

This second count is verified by a bipartite matching algorithm applied to a special bipartite graph with vertices: k disjoint copies of V on one side and E and the tie-down edges T_i on the other. The edges join any edge (or tie-down) to all copies of adjacent vertices. This matching is examined in more detail in Tay [22].

Remark 2.23. If we think of the tied-down graph (with tie-down edges as loops), the bipartite matching of Remark 2.22 actually covers the graph G_i with k edge-disjoint independent sets in the bicircular matroid of the graph. These sets replace the trees of Theorem 2.18, which are independent sets of the cycle matroid of the graph. A tree with a loop at the root is one example of such a set, but there are others (Fig. 2.3A, B). (A minimal dependent set of the bicircular matroid is a "bicycle" (Fig. 2.3C).)

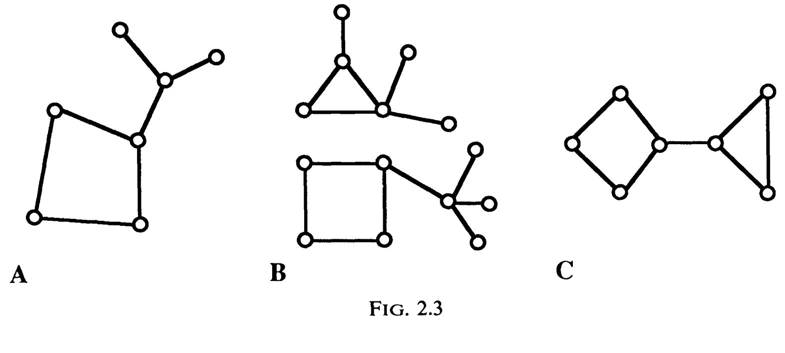


FIG. 2.3

Remark 2.24. For any vertex i, the bipartite matching associates k adjacent edges to each vertex (and associates the tie-down edges with the initial vertex i). This gives a k-fan for the graph, and gives a starting point for implementing the Rosenberg-type method, without finding any trees. Of course the existence of such a k-fan for one initial vertex is not sufficient to show that the graph is k-isostatic. The existence of a k-fan for each choice of initial vertex is both necessary and sufficient for the graph to be k-isostatic, by Remark 2.22.

3. Motions and examples.

3.1. Examples of pure conditions. Let us consider some more examples of isostatic bar-and-body frameworks, frames, and their pure conditions. We begin with some planar frameworks, which are examples of 3-frames.

Example 3.1. (Fig. 3.1.) Recall that a is the 2-extensor determined by the line in which the corresponding bar lies, and that in the plane a is itself a 3-tuple. I_3 is the 3×3 identity. We see that we may directly expand the 9×9 determinant det $M(G, T)$ as

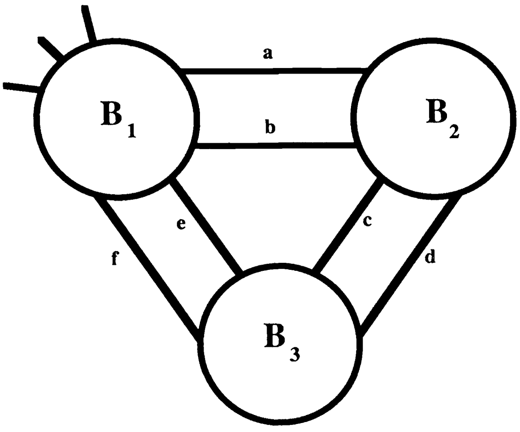
$$C(G)=[abd][cef]-[abc][def],$$

where brackets denote ordinary 3×3 determinants. This may also be obtained from the two 3-fans (abd), (cef) and (abc), (def).

This pure condition has an interpretation in the dual Cayley algebra (see [5]) as follows: rewriting $C(G)=(a\wedge b)\vee(c\wedge d)\vee(e\wedge f)$, we see that $C(G)=0$ precisely when the three points of intersection determined by the pairs of bars are collinear. This result illustrates a thorem of Arnhold and Kempe:

> If three bodies are in motion in the plane, the relative centers of motion of the three pairs of bodies are collinear.

We will see shortly that (a scalar multiple of) ab represents the relative center of motion of the bodies B_1 and B_2 if there is a motion.



$$M(G,T) = \left(\begin{array}{ccc} B_1 & B_2 & B_3 \\ a & -a & \\ b & -b & \\ & c & -c \\ & d & -d \\ e & & -e \\ f & & -f \\ I_3 & & \end{array}\right)$$

FIG. 3.1

We also remark that in this example we could have tied down another body, say B_2. Then we would obtain $C(G)=[abe][cdf]-[abf][cde]$. Although this appears to be different from the previous bracket expression for $C(G)$, it can be shown by the use of syzygies (i.e. standard determinantal identities) that the two are equal, illustrating Proposition 2.8. The equality of these expressions may also be inferred from the symmetry of the dual Cayley algebra expression given above.

Example 3.2. Next we consider the framework in space shown in Fig. 3.2A. Using the three 6-fans shown in Fig. 3.2B, we have:

$$C(G) = +[abcdej][fghikl] - [abcdek][fghijl] + [abcdel][fghijk].$$

This also has a Cayley algebra factoring as $C(G) = (abcde) \wedge (jkl) \wedge (fghi)$ which may be interpreted geometrically as follows. There is a one-dimensional space of relative motions between B_1 and B_2. Similarly there is a two-dimensional space of relative motions between B_1 and B_3 and a three-dimensional space between B_2 and B_3. Looking at all three of these subspaces of the six-dimensional space of all possible motions, $C(G) = 0$ says that the three subspaces are not independent, i.e. some relative motion between B_1 and B_3 is a linear combination of a motion between B_1 and B_2 and a motion between B_2 and B_3.

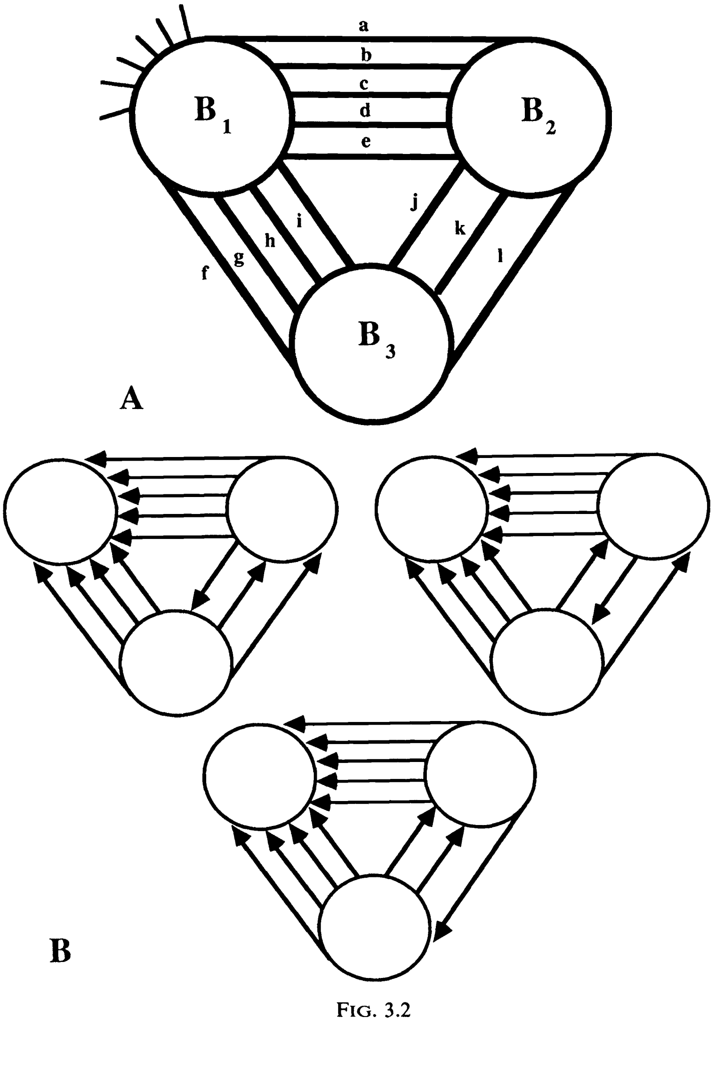


FIG. 3.2

3.2. Motions of 1-underbraced frameworks. Let us now consider a k-frame which is 1-*underbraced*, that is, one edge short of a k-counted graph. Such a k-frame will generically have one k-motion besides the trivial ones. We now develop an algebraic method of describing this motion. More specifically, we will think of vertex number one as tied down, and compute the relative motions of the other vertices relative to vertex one.

Let us consider the k-frame matrix $M(G(p), T)$ for such a k-frame.

$$\begin{array}{c} \begin{array}{ccc} V_1 & V_2 & \cdots \end{array} \\ \begin{bmatrix} a & -a & 0 \\ b & 0 & -b \\ \cdots & \cdots & \cdots \\ I_k & 0 & 0 \end{bmatrix} \end{array}.$$

Assuming that the $kv-1$ rows of $\mathrm{sp}\,(M)$ are linearly independent, there is a unique (up to scalar) row vector Z^* of length kv orthogonal to the rows of M. We may compute Z^* by adding a row of indeterminates, $(x_1, x_2, \cdots, x_{kv})$, to M, and setting the determinant equal to zero. The coefficient of x_i in the resulting equation is $(-1)^i$ times the ith component of Z^*. This may be checked by elementary linear algebra and Cramer's Rule.

Let vertex V_i undergo a motion Z_i, meaning that Z_i is the k-vector of components of Z^* in the columns for V_i. Thus Z_i is the vector of coefficients of the k indeterminates in the columns for V_i. These k indeterminates may be regarded as an indeterminate vector V_e for a new edge between V_1 and V_i.

THEOREM 3.3. *In a 1-underbraced k-frame G, with V_1 tied down, the motion of a vertex V_i may be computed by adding a dummy edge $x=(x_1, x_2, \cdots, x_k)$ between V_1 and V_i, computing the pure condition of $G \cup x$, and taking the coefficients of the component of x.*

COROLLARY 3.4. *The relative motion of a vertex V_j with respect to a vertex V_i is obtained by adding a dummy edge x between V_i and V_j and computing the coefficients of x in the pure condition of $G \cup x$.*

COROLLARY 3.5. *The relative motion between V_i and V_j may be expressed in terms of the Cayley Algebra as the $(k-1)$-extensor obtained from the pure condition of $G \cup x$, where x is a dummy edge between V_i and V_j, by deleting x from the bracket expression for the pure condition of $G \cup x$.*

3.3. Examples of motions.

Example 3.6. (Fig. 3.3.) In this 1-underbraced 3-frame, the relative motion of B_2 with respect to B_1 is obtained by adding an edge x between B_1 and B_2, obtaining the pure condition $[abx][cef]$. Thus the relative motion is $[cef]ab$. Regarding this 3-frame as a bar-and-body framework in the plane, the join ab of the vectors a and b is the same as the meet $a \wedge b$ with a and b regarded as 2-extensors (or, in this case, co-vectors). In Example 3.1, when the pure condition $C(G)$ is zero, and the framework has a single motion, any one of the six bars is dependent upon the other five; hence the motion of the framework is the same as that of the 1-underbraced framework of this example. Thus we have justified the statement in Example 3.1 that a scalar multiple of $a \wedge b$ is the center of the relative motion of B_1 and B_2. Geometrically, the center of the motion is the point of intersection of the lines a and b, regardless of the scalar multiple. The scalar, $[cef]$, affects only the velocity of the rotation of B_2 about the point $a \wedge b$, but this is important in comparing or combining the relative motions of several bodies in the same framework.

The relative motion of B_3 relative to B_1 may similarly be computed as $[abc]ef$, and the motion of B_3 relative to B_2 as $[abc]ef-[cef]ab$ (using the pure condition computed in Example 3.1), and as we certainly expected, relative motions are additive:

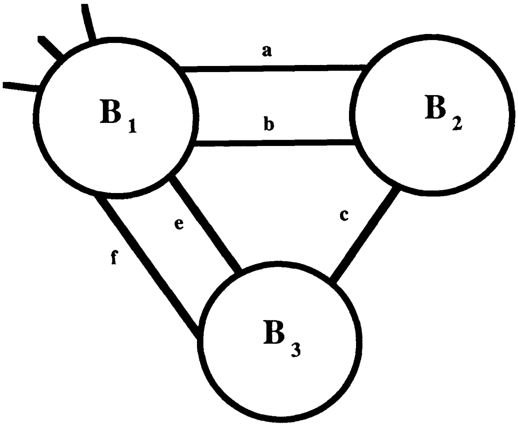


FIG. 3.3

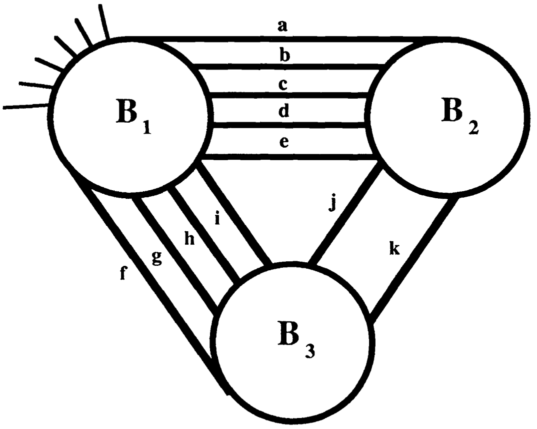
if M_{ij} is the relative motion of B_i relative to B_j then $M_{ij}=M_{ih}+M_{hj}$ for all h, and $M_{ij}=-M_{ji}$.

Example 3.7. This example is a 1-underbraced bar-and-body framework in three-dimensional space; hence a 6-frame (Fig. 3.4).

Here M is a 17×18 matrix. Applying Corollary 3.5,

$$Z_1=0,$$
$$Z_2=[fghijk]abcde,$$
$$Z_3=[abcdek]fghij-[abcdej]fghik$$

where we have used 6-fans with the obvious conventions to compute Z_3.



$$M(G,T)=\begin{bmatrix} a & -a & \\ b & -b & \\ c & -c & \\ d & -d & \\ e & -e & \\ f & & -f \\ g & & -g \\ h & & -h \\ i & & -i \\ & j & -j \\ & k & -k \\ I_6 & & \end{bmatrix}$$

FIG. 3.4

An expression such as *abcde* in Z_2 is a join of five "vectors" in the vector space $V^{(2)}$ of 2-extensors, a six-dimensional vector space. Such a join corresponds to a five-dimensional subspace U of $V^{(2)}$, which corresponds to a line complex (see [4], [10]), generated by five lines. In fact, an arbitrary five-dimensional subspace U corresponds to a line complex, sometimes called a linear line complex. The line complex is actually the set of vectors in U which correspond to lines. A line complex has a unique line or screw reciprocal to it, and for *abcde*, that line or screw is the center Z_2 of the motion of B_2, up to scalar multiple. A line is reciprocal to Z_2 precisely when a bar on that line does not (instantaneously) block the motion with center Z_2. Thus if Z_2 is a line, the line complex consists of all lines meeting Z_2, whereas if Z_2 is a screw, every line of the complex is, at each of its points, normal to the velocity vector of the screw motion. If *ab* is a line, then *ab* is in the complex corresponding to Z_2 if and only if $Z_2 \vee ab = 0$, agreeing with Proposition 1.3.

We have a linear combination of two such expressions for Z_3^*, but we may also write Z_3^* in the factored form $fghi(\alpha k - \beta j)$, where the last factor, a linear combination of the lines k and j, is, in general, not a line. Nevertheless, Z_3 is geometrically either a line or a screw, as every instantaneous motion of B_3 is a translation, a rotation, or a screw motion.

3.4. Motions in special position. Now let us consider an overbraced k-frame $G(p)$, that is, one in which the rows of $M(G)$ are dependent. If we select a basis of the row-space of $M(G)$ and delete all other edges of G, then we have clearly not changed the motion space of $G(p)$. This fact must be reflected in our Cayley algebra calculations.

Example 3.8. We return to the 3-frame in Example 3.1 (see Fig. 3.1). Now, however, we assume that the 3-frame is in the special position

$$(*) \qquad 0 = C(G) = (a \wedge b) \vee (c \wedge d) \vee (e \wedge f) = [abc][def] - [abd][cef].$$

Thus the frame, though correctly counted, is underbraced, and in a generic point for this special position [see § 4], the six edges form a circuit. Thus any one of the six edges may be removed without affecting the single motion of $G(p)$. Let $Z_2^{(u)}$ denote the center of the relative motion of B_2 to B_1, computed by removing edge U. Then

$$Z_2^{(u)} = [cde]bf - [cdf]be.$$

Similarly

$$Z_2^{(v)} = [cde]af - [cdf]ae, \qquad Z_2^{(w)} = [def]ab,$$

etc.

Now $Z_2^{(u)}$ and $Z_2^{(v)}$ must represent the same geometric motion, and hence must be scalar multiples of each other in the Cayley algebra. Indeed $[acd]Z_2^{(u)} - [bcd]Z_2^{(v)} = 0$ as may be verified using standard syzygies and (*). Similarly

$$[abd]Z_2^{(u)} + [cbd]Z_2^{(w)} = 0.$$

Some care must be taken in choosing the coefficients, in addition to assuring homogeneity and the correct sign. For example,

$$[aef]Z_2^{(u)} \pm [cef]Z_2^{(w)} \neq 0.$$

For a more detailed examination of relative motion in special positions, see § 4.3.

4. Factors, motions and stresses. We have seen that deleting an edge from a k-isostatic frame gives a single internal motion. The same motion occurred if this edge was reinserted in an appropriate special position such that the coordinates satisfied the pure condition.

In this section we will give a description of the broad pattern of this motion and the form of these special positions, based entirely on the factors of the pure condition and on simple combinatorial properties of the graph.

4.1. The lattice of blocks. The basic units for our study will be the subgraphs of a graph which are themselves k-isostatic, and the irreducible factors of the pure condition. From previous work on bar-and-joint frameworks, we anticipate that some of the factoring of the pure condition is explained by the presence of such k-isostatic subgraphs [25, Prop. 4.4]. In fact, we will show that all the factoring and other associated properties arise from this source.

With this goal in mind, we begin with a basic result about these subgraphs.

DEFINITION 4.1. A *block* of a k-isostatic graph G is a subgraph G' which is also k-isostatic.

We observe that the block G' gives a block decomposition of $\det(M(G, T))$ provided the tie-down T is attached to a body of G' (see the proof of Proposition 4.4).

THEOREM 4.2. *The blocks of a k-isostatic graph G, ordered by inclusion, form a lattice, with* $G_1 \wedge G_2 = G_1 \cap G_2$ *for any two blocks.*

Proof. (a) Given any two blocks, G_1 and G_2, we will show that $G_1 \cap G_2$ is a block. If $G_1 \cap G_2$ is empty, then it is, in a trivial way, k-isostatic.

If $G_1 \cap G_2 = G_3$ is nonempty, then we know that this subgraph of G satisfies $|E_3| \leqq k(|V_3|-1)$. We set $m = k(|V_3|-1)-|E_3|$, and show that m is also $\leqq 0$.

Since G_1 and G_2 are nonempty blocks, we know that $|E_1| = k(|V_1|-1)$ and $|E_2| = k(|V_2|-1)$. Consider the graph $G' = G_2 \cup G_2$. By the inclusion-exclusion principle:

$$|V'| = |V_1| + |V_2| - |V_3|$$

and

$$|E'| = |E_1| + |E_2| - |E_3| = k(|V_1|-1) + k(|V_2|-1) - (k(|V_3|-1) - m)$$
$$= k(|V_1| + |V_2| - |V_3|) - k + m = k|V'| - k + m.$$

Since G' is contained in G, we know that $|E'| \leqq k(|V'|-1)$ and $m \leqq 0$. Thus $|E_3| = k(|V_3|-1)$ and G_3 is a block.

(b) The partially ordered set of blocks is a finite set with maximum element G and minimum element ϕ. For any two blocks G_1 and G_2 we define:

$$G_1 \vee G_2 = \bigcap G' \qquad \text{(intersection over blocks } G' \supset G_1,\ G' \supset G_2).$$

By part (a) this nonempty intersection is a block—the unique minimum block containing G_1 and G_2. Thus, by a standard construction we have a lattice. □

Remark 4.3. If $G_1 \cap G_2 \neq \phi$, then we find, from the count in part (a), that $G_1 \cup G_2$ is a block. In this case $G_1 \vee G_2 = G_1 \cup G_2$.

In Fig. 4.1A we show a 3-isostatic graph and its associated lattice. (A block is described by listing its vertices.) In Fig. 4.1B we show a 2-isostatic example with its lattice.

This lattice of blocks gives a partition of the edges of G. For each edge e (joining vertices B and C) there is a lowest block G_e which contains the edge ($G_e = B \vee C$). We say that e is *associated with* G_e. Figures 4.2A and B illustrate the associated edges for the lattices of Figs. 4.1A and B.

The lattice of blocks also partitions the factors of the pure condition of G. We recall that an *irreducible polynomial* over a field K is a polynomial f (with at least one variable) such that, if $f = g \cdot h$ over the field then either g or h is the zero element of the field. Since the pure condition of a k-isostatic graph is a polynomial over the

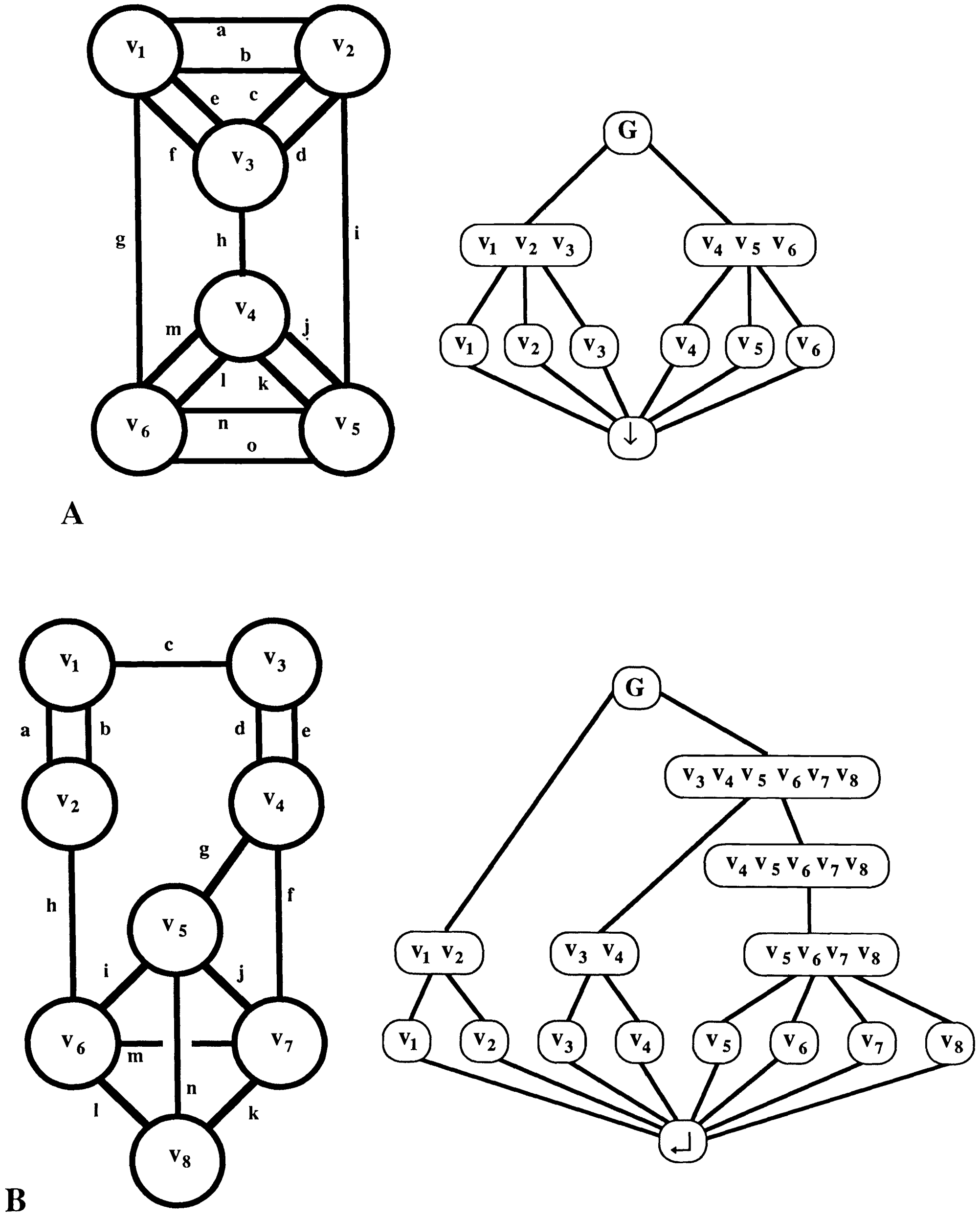

A
v1
v2
v3
v4
v5
v6
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
G
v1 v2 v3
v4 v5 v6
v1
v2
v3
v4
v5
v6
↓
B
v1
v2
v3
v4
v5
v6
v7
v8
a
b
c
d
e
f
g
h
i
j
k
l
m
n
G
v3 v4 v5 v6 v7 v8
v4 v5 v6 v7 v8
v1 v2
v3 v4
v5 v6 v7 v8
v1
v2
v3
v4
v5
v6
v7
v8
↵

FIG. 4.1

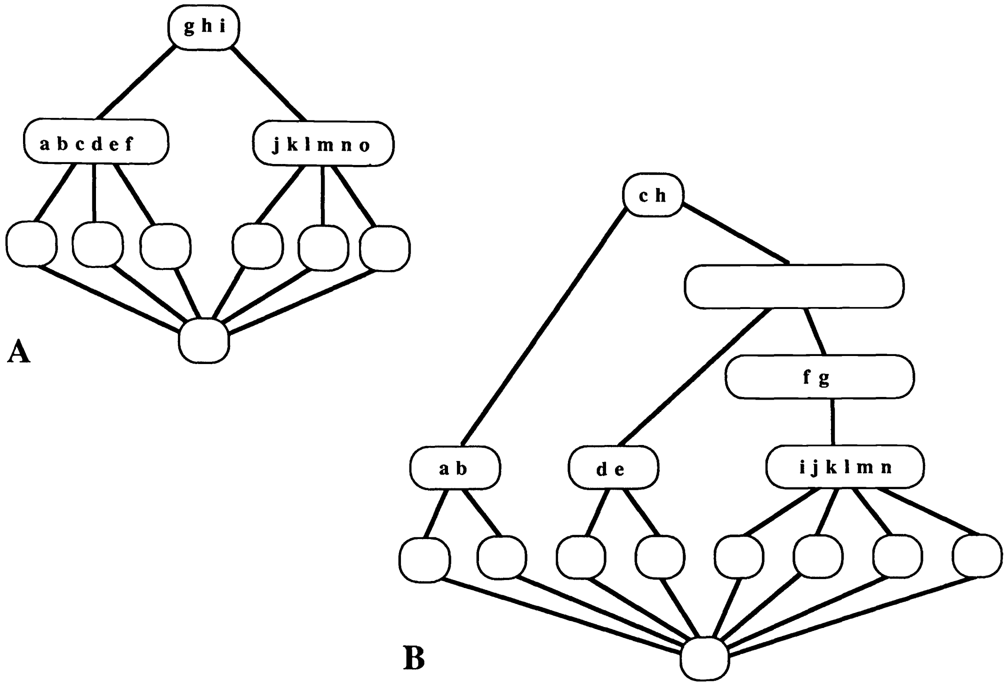


FIG. 4.2

rationals which is of first degree in every variable, it is a simple result of commutative algebra that any irreducible factor over the rationals is also irreducible over the complex numbers (or any other field extending the rationals). For simplicity we speak of an *irreducible factor of the k-isostatic graph G.*

PROPOSITION 4.4. *Each irreducible factor f of a k-isostatic graph G is associated with a unique block G_f such that*:

(i) *f is a factor of the pure condition of G_f, and G_f is minimal among such blocks*;

(ii) *each edge with variables occurring in f is also associated with the block G_f.*

Proof. (a) Any block G' has $|E'| = k(|V'|-1)$. We reorder the rows and columns of the k-frame matrix of G so that the vertices and edges of G' come to the upper left corner. The matrix now looks like

$$M(G) = \begin{bmatrix} C(G') & 0 \\ H & L \end{bmatrix}.$$

When we tie down the first vertex and take determinants, we find that

$$C(G) = C(G') \cdot \det(L).$$

Thus $C(G')$ is a factor of $C(G)$. By the form of L it is clear that all variables for edges in G' occur only in $C(G')$.

(b) Any irreducible factor of G is either of first degree in each variable of an edge or has no occurrences of variables for that edge. This follows from the fact that $C(G)$ is linear in the vector for the edge, and any factoring must preserve such homogeneity [25, Thm. 2.1].

(c) Each edge occurs in a unique lowest block G_e, and the variables for the edge occur in a unique irreducible factor f_e. Therefore f_e must be a factor of G_e. However, if a block G' does not contain the edge e, then f_e cannot be a factor of $C(G')$. We conclude that the factor f is associated with the block G_e. Clearly this assignment would be the same for every edge occurring in the irreducible factor and we have defined the associated block G_f. □

When some variables for an edge (and therefore all variables for the edge) occur in an irreducible factor f, we say that *the edge occurs in f.*

In Figs. 4.3A and B, we give the associated factors in the blocks for the examples of Fig. 4.1. This lattice of blocks, with its associated factors, gives a basic outline of the structure of G. For any block G' of G, the lattice of blocks of G' is just the sublattice from ϕ to G' defined inside G. The pure condition for each block G' will be the product of all irreducible factors for blocks $\leqq G'$ in the lattice.

4.2. The scope of an irreducible factor. If the vectors for the edges of a k-isostatic graph are specialized so that the pure condition is 0, then the k-frame will have an internal motion. This drop in the column rank of the k-frame matrix must correspond to a drop in the row rank—a row dependence. Such a row dependence is analogous to a static stress in a bar and joint framework [25, § 1].

DEFINITION 4.5. A *stress* on a k-frame $G(p)$ is an assignment of scalars $\lambda(e_i)$ to the edges of the graph, such that $(\cdots, \lambda(e_i), \cdots)$ is a row dependence of the k-frame matrix for $G(p)$.

The *scope* of a stress is the set of edges of G which have nonzero scalars in the stress.

We say a frame is *independent* if it has only the trivial stress with all scalars zero. Thus, for example, an independent k-frame with $|E| = k(|V| - 1)$ will be isostatic, since it has the required row (and column) rank.

The set of stresses on a k-frame is a vector space—the space of stresses. As a convenient short hand, when this space has dimension 1, we speak of a *single stress.*

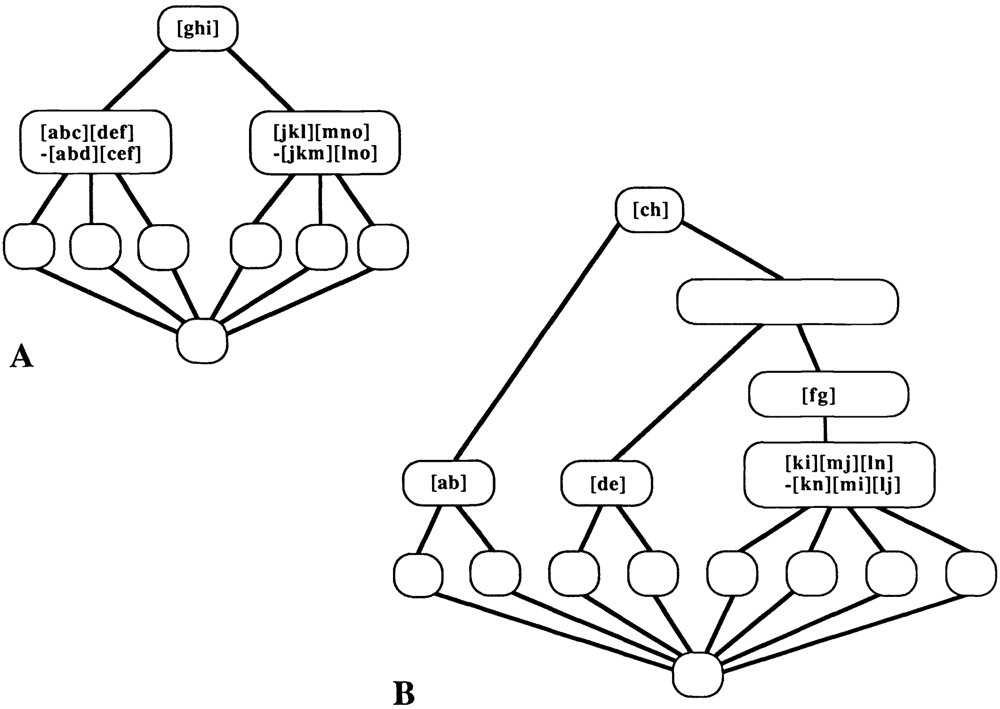


FIG. 4.3

For any irreducible factor f of a k-isostatic graph G, we can find a generic point p of f: values for the variables such that $f(p)=0$ and such that, whenever $g(p)=0$ for another polynomial, then $g=f\cdot h$ [8, pp. 10–15], [31, p. 21]. For a general irreducible polynomial, the generic point lies in some general extension of the field. Since an irreducible factor of any pure condition is a polynomial over the rationals of first degree in each variable, we can always find generic points in R^n. If circumstances dictate the addition or deletion of variables which do not occur in f, we can simply delete, or add corresponding algebraically independent real numbers, without changing the generic character of the point. In this chapter, such changes to the point will pass without further comment. We also note that if the value for at least one variable in f is omitted, the remaining coordinates are algebraically independent.

PROPOSITION 4.6. *A generic point p of an irreducible factor for the k-isostatic graph G defines a single stress on the k-frame $G(p)$, whose scope is the entire block G_f.*

Proof. (a) If we add an extra edge d to the graph G_f, joining some pair of vertices also joined by an edge e of f, then we create a graph G' whose general position frame has a single row dependence. Using Cramer's rule on the tied-down k-frame matrix for G', and writing $R(e_i)$ for the row of the edge e_i, this row dependence can be written

$$\Sigma \pm C(G'-e_i)\cdot R(e_i)+C(G_f)\cdot R(d)=0.$$

(See [25, § 5] for details of the similar expansion on a framework.) When we specialize to a generic point p of f, $C(G_f(p))=0$ and the equation becomes

$$(1) \qquad \Sigma \pm (C(G'-e_i)(p))\cdot R(e_i(p))=0.$$

Since $G'-e$ is isomorphic to G_f, with d replacing e, we know that $C(G'-e)\neq 0$. Since f cannot be a factor of this polynomial (different variables), and p is a generic point of f, we know that $(C(G'-e)(p))\neq 0$. Equation (1) expresses a stress on edges in G_f. Since removing the edge e from G' creates an isostatic k-frame, we know there is exactly one stress on $G'(p)$, and therefore on $G_f(p)$. We also know that the edge e is in the scope of the stress, and this must be true for every edge occurring in f.

(b) Take a point q making $G_f(q)$ isostatic and a generic point p of f. We select an edge e occurring in f, joining vertices B and C, and tie down vertex B. Assume there is an edge d in G_f which is not in the scope of the stress on $G_f(p)$, and therefore not in f, and remove this edge to create G'. We will derive a contradiction.

Assume C does not move in $G'(q)$. Then B and C must lie in a k-isostatic subgraph of G'. This contradicts the minimality of G_f among blocks containing e.

Assume C does move in $G'(q)$. Inserting a new edge e', with general coordinates, also joining B and C, creates a new k-isostatic graph G_1, since it blocks this motion. However $G_1(p)$ must contain a stress, since it includes the scope of the stress in $G_f(p)$, by assumption. Therefore, $C(G_1(p))=0$ for a generic point of f, and $C(G_1(p))=fh$. Since e and e' join the same two vertices, setting the variables for e and e' the same will always create a row dependence, and make $C(G')=0$. Therefore the two sets of variables must both occur in some irreducible factor of $C(G')$. However the variables of e occur only in f, and no variables of e' occur in f (since e' was not in G_f). We have reached the desired contradiction. □

Figure 4.4A shows a special position of the graph of Fig. 4.1A, with the three lines g, h, i concurrent. This is a generic special position for the factor $[ghi]$. Since all edges lie below this factor in the lattice (Fig. 4.4B), this position defines a stress whose scope is the entire graph G (shown by arrows on the edges). Figure 4.4C shows a special position for the factor $[fg]$ in the graph of Fig. 4.1B. For this position the scope of the stress is the block shown in Fig. 4.4D.

Since every generic point of a factor *f* creates a stress with the same scope, we speak of the *scope of the irreducible factor.*

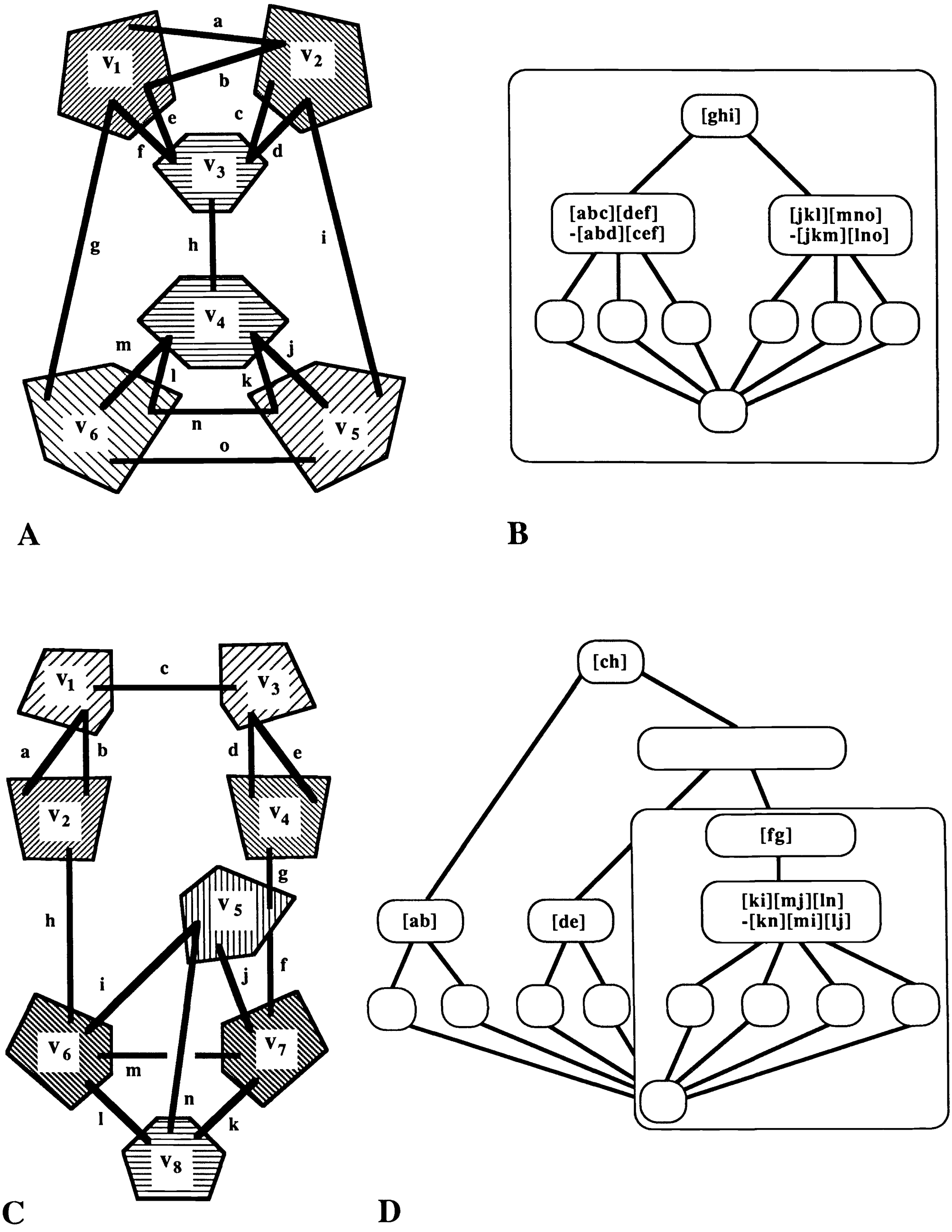


FIG. 4.4

4.3. The motion of an irreducible factor. For any generic point p of an irreducible factor f, we know that $G(p)$ has a single stress. If we tie down the first vertex, this will leave a single motion (up to scalar multiplication). We want to investigate which sections of the frame are locked together, and which pairs of vertices go into relative motion.

DEFINITION 4.7. A *component* of a motion M of a frame $G(p)$ is a maximal subgraph G' such that no two vertices in G' are in relative motion in M. A *link* of the motion is an edge of G which joins vertices which are in relative motion.

PROPOSITION 4.8. *For any generic point p of the irreducible factor f of the k-isostatic graph G, the components of the motion of $G(p)$ are the maximal blocks of G which do not contain G_f.*

Proof. (a) Consider a block G, not containing G_f. At any generic point p of f we can delete an edge occurring in f and leave an independent k-frame with the same motion as $G(p)$. Therefore $G_1(p)$ is an independent k-frame, with $|E_1| = k(|V_1| - 1)$. It is isostatic and must lie inside a single component of the motion.

(b) Consider a component G_2 of the motion of $G(p)$. Once more we delete an edge e occurring in f, to create an independent k-frame G'. Any pair of vertices, A and B, of G_2, must share a k-isostatic subframe $G(A, B)(p)$ in G'. For any three vertices, $G(A, B)$ and $G(B, C)$ are isostatic subframes sharing at least a vertex, so their union is an isostatic subframe (see Remark 4.3). By induction on the vertices, we see that G_2 is contained in a k-isostatic subgraph G_3 of G'. By part (a) G_3 is contained in a component, so $G_2 = G_3$ and the component is a block not containing G_f. □

In Figs. 4.5A and B we show the components (lightly outlined boxes) and links (heavy lines) of the special positions of Fig. 4.4A and C.

Since all generic points of an irreducible factor f create the same components (and by subtraction, the same links), we speak of the *components and links of the irreducible factor.*

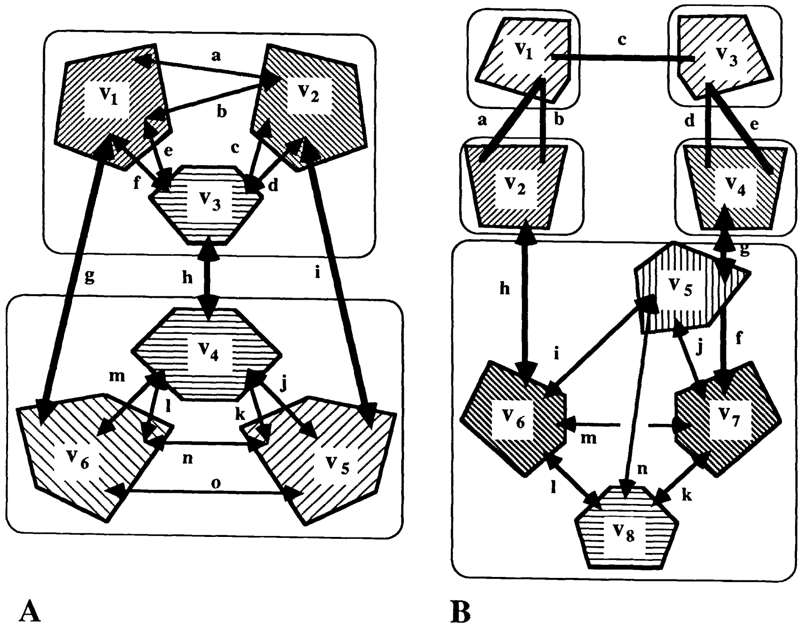


FIG. 4.5

THEOREM 4.9. *An edge occurs in the irreducible factor f of a k-isostatic graph if and only if the edge is in the scope of f and is a link of f.*

Proof. (a) Assume the edge occurs in f. The edge lies in G_f, so by Proposition 4.6 it lies in the scope of f. The edge does not lie in any block not containing G_f. By Proposition 4.8, the edge does not lie in component—so it must be a link.

(b) Assume an edge b is in the scope of f and in the link of f. At any generic point p of f, the removal of the edge b leaves a k-frame with no stress, and one motion (after a tie-down). Since b was in the link, the vertices joined by b are in relative motion. Inserting a general vector for b will block this motion, and create a point q with $f(q) \neq 0$. We conclude that the variables for b actually occur in f. □

The picture of how irreducible factors are associated to the lattice of blocks can now be completed.

For the examples of Figs. 4.4 and 4.5, we see that the edges f, g, h (resp. the edges f, g) are in both the scope of the factor and the links of the factor.

THEOREM 4.10. *At most one irreducible factor is associated with a block of a k-isostatic graph. All edges associated with the block occur in this factor.*

Proof. Consider the block G_f for some irreducible factor. All edges of G_f are in the scope of f. If an edge b is also a link of f, then, by Theorem 4.9 it occurs in f. Otherwise b is in a component of f, a block G' not containing G_f. Therefore b is in the block $G' \cap G_f$ below G_f. The edge, and any related factor, is not associated with G_f. □

4.4. The reduced graph of a factor. Each generic point of an irreducible factor has an associated internal motion and stress. In fact, from the lattice of blocks, with the associated edges, and the position of the block G_f, we can identify the scope of the stress, and the components and links of the internal motion. The essential features of this factor will be even clearer if we shrink each component of the motion to a single vertex, and restrict our attention to the scope of f.

DEFINITION 4.11. For an irreducible factor f of a k-isostatic graph G, the *reduced graph of the factor* is the graph formed from G_f by contracting all edges of blocks below G_f.

THEOREM 4.12. *The reduced graph of an irreducible factor of a k-isostatic graph G is a k-isostatic graph, with edges in $1-1$ correspondence with the edges occurring in f. The pure condition of the reduced graph is f (up to a scalar).*

Proof. Each edge of G_f is either in a component of f, and thus contracted out, or is a link, and appears in the reduced graph. By Theorem 4.9 such an edge occurs in f. Conversely all edges occurring in f are links in G_f, and survive to form edges of the reduced graph.

Given a motion of the reduced graph, there is a corresponding motion of G_f which transfers the motion of the vertex to the associated block G_f. Since, in general position, G_f has only trivial motions, the reduced graph is also rigid in general position. If we delete any edge of the reduced graph then we get the motion corresponding to deleting an edge occurring in f from G_f. Thus deletion of any edge of the reduced graph, in general position, causes an internal motion. We conclude that a general position realization as a k-frame is minimal and rigid, so the reduced graph is k-isostatic. At a generic point p of f, the frame $G_f(p)$ has an internal motion in which blocks below G_f are components. The reduced graph will have the corresponding internal motion, and p must make its pure-condition zero. Therefore f divides the pure condition of the reduced graph. Since f incorporates all possible occurrences of variables for edges of the reduced graph, it is, up to a scalar, the pure condition. □

Figure 4.6A shows the reduced graph of the factor $[ghi]$ in the example of Fig. 4.1A, while Fig. 4.6B shows the reduced graph of $[fg]$ in Fig. 4.1B.

Remark 4.13. The reduced graph of a factor has a lattice of blocks which is very simple: a bottom—ϕ, a top—the graph, and a set of pairwise incomparable middle points—one for each vertex. Conversely, any k-isostatic graph with such a lattice cannot be further reduced, so it must have an irreducible polynomial as its pure condition.

Such a k-isostatic graph G with an irreducible pure condition is also recognizable by the simple counting property: $|E|=k(|V|-1)$ and $|E'|<k(|V'|-1)$ for all proper subgraphs. Such graphs are *k-irreducible.*

A k-irreducible graph shows a striking uniformity of both static and kinematic behavior as a general position k-frame.

If we delete an edge from a generic, k-isostatic realization of the graph, the resulting motion puts all pairs of vertices into relative motion. (This motion is equivalent to the motion at a generic point of the pure condition, with the deleted edge in special position). This is a *complete motion,* a concept which has applications in scene analysis (see § 5).

If we add an extra edge d, in general position, to a generic isostatic realization of the k-irreducible graph, $G(p)$, the resulting stress will have the entire extended graph as its scope. (Deletion of any other edge from the extended k-frame is equivalent to deleting an edge of $G(p)$ and inserting a general edge between vertices which are in relative motion. Therefore no subgraph contains the stress.) Such a graph is a general position *circuit,* a concept which is common in matroid theory and in the study of tensegrity frameworks [17].

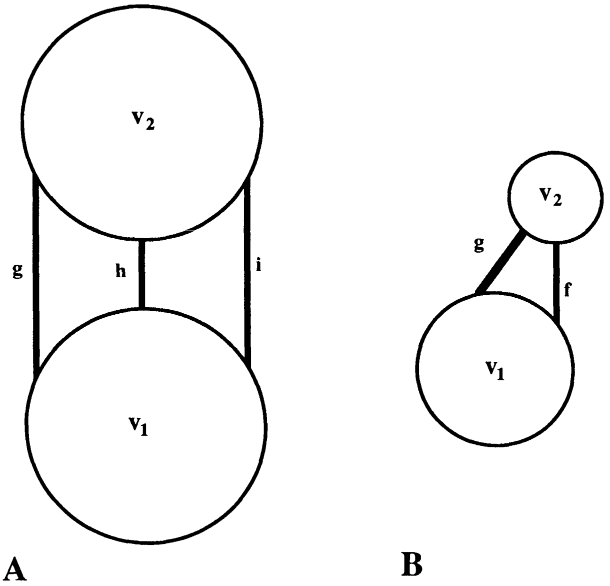


FIG. 4.6

4.5. Generic points of several irreducible factors. We know that distinct irreducible factors actually use disjoint sets of variables.

DEFINITION 4.14. A *generic point* of the set $\{f_i\}$ of irreducible factors of a k-isostatic graph G is a point p in R^{kv} such that for each i, when p is restricted to variables not in $f_j, j\neq i$, p is a generic point of f_i.

Remark 4.15. If the value of at least one variable in f_i is omitted for each i, the remaining coordinates of a generic point will be algebraically independent over the rationals.

Remark 4.16. In algebraic geometry, generic points are defined for irreducible varieties [8, pp. 10–15]. The set of irreducible factors $\{f_i\}$ defines a variety

$$V(\{f_i\})=\{(x_1,\cdots,x_m)\in R^m|f_i(x)=0\ \forall i\}.$$

Since the f_i are irreducible, with disjoint sets of variables, this is indeed an irreducible variety. Its generic points are defined as points p such that $g(p)=0\rightarrow g(X)=0$ for all X in $V(\{f_i\})$. In fact it can be shown that these are identical to the generic points defined above. We have chosen to emphasize the property of generic points which will be used in our proofs.

THEOREM 4.17. *Assume p is a generic point of $\{f, g\}$ for two irreducible factors of the k-isostatic graph G.*

(i) *If G_f and G_g are incomparable in the lattice of blocks, then $G(p)$ has a* 2-dim *space of stresses generated by the stresses of f and g, and a* 2-dim *space of internal motions, generated by the motions of f and g.*

(ii) *If $G_f<G_g$, then $G(p)$ has the single stress of f and the single internal motion of g.*

Proof. (i) Assume G_f and G_g are incomparable. If we restrict to G_f, this omits all edges of g and leaves a generic point of f. We have the stress of f in G_f and in G. Similarly we have the stress of g and these stresses (with different scopes) are independent. The dimension of the space of stresses is at least 2.

The space of internal motions is now at least of dimension 2. If we delete one edge occurring in f, and one occurring in g, this leaves an independent k-frame with $|E'|=k(|V'|-1)-2$. Therefore the space of motions has dimension exactly 2. If we reinsert the edge from g, with general coordinates, this leaves only the motion of f, so this motion appears in $G(p)$ as well. Similarly, the motion of g also occurs in $G(p)$. These independent internal motions of f and g generate the entire space of internal motions (modulo a tie-down of the first vertex).

This also shows that the space of stress of $G(p)$ has dimension 2, so it is generated by the stresses of f and g.

(ii) Assume $G_f<G_g$. If we restrict to the graph G_f, we omit all edges of g and have a generic point of f. Therefore there is at least the stress of f, with scope G_f, in $G(p)$. If we delete an edge occurring in f, $G_f(p)$ becomes an independent, general position k-frame. In $G_g(p)$, the omitted edge would be part of the scope of g (if placed in general position). Therefore $G_g(p)$, with this edge deleted is an independent k-frame with remaining coordinates for a generic point of g. It has the single internal motion of g (the same motion which occurs with the edge inserted in general position). We conclude there is exactly the single stress of f and the single motion of g. □

Figure 4.7A shows a special position for the graph of Fig. 4.1B which is a generic point of the incomparable factors $[fg]$ and $[ab]$. The scopes of the two stresses are shown with arrows, and the links and components of a general combination of the two motions are drawn with heavy lines and light boxes, as before. Figure 4.7B illustrates a special position for the graph of Fig. 4.1A which is a generic point for the two factors $[abc][def]-[abd][cef]<[ghi]$. The scope of the single stress and the links and components of the single motion are also shown in the figure.

This analysis can be extended to describe the patterns of motions and stresses at a generic point p of three factors f, g, h of a k-isostatic graph G. We offer, without proof, a summary of the cases:

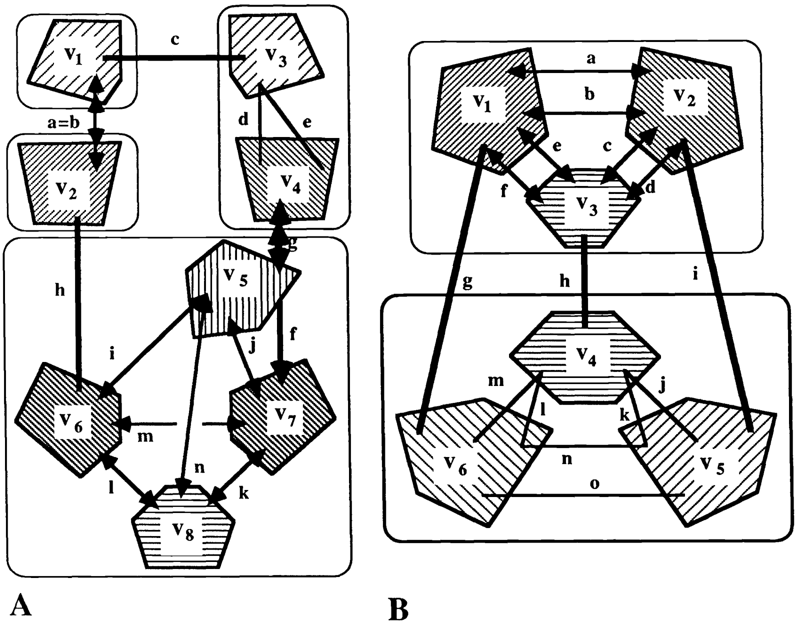


FIG. 4.7

(i) If no two blocks G_f, G_g, G_h are comparable, the $G(p)$ has a 3-dim stress space generated by the stresses of f, g and h, and a 3-dim space of internal motions generated by the motions of f, g and h.

(ii) If $G_f < G_g < G_h$, then $G(p)$ has the single stress of f and the single internal motion of h.

(iii) If $G_f < G_g$, G_h and G_g, G_h are incomparable, then $G(p)$ has a 2-dim space of motions generated by the motions of g and h, and a 2-dim stress space including the stress of f.

(iv) If G_f, $G_g < G_h$ and G_f, G_g are incomparable, then $G(p)$ has a 2-dim space of stresses generated by the stresses of f and g, and a 2-dim space of internal motions which includes the motion of h.

For a generic point p of a general set of irreducible factors $f, g, \cdots, h$, we know that:

(i) If G_f is minimal among blocks for the set, then the stress of f is in the stress space of $G(p)$.

(ii) If G_g is maximal among blocks for the set, then the motion of g is in the space of internal motions of $G(p)$.

(iii) If there is a set of m pairwise incomparable blocks for the set, then the space of stresses (and of internal motions) has dimension at least m.

We *conjecture* that:

(iv) The dimension of the space of stresses (and the space of internal motions) is exactly the size of the largest set of pairwise incomparable blocks for the set of factors.

4.6. Finding graphs for given factors. For a mathematician, it is natural to ask whether a pure condition arises from a unique graph, or a unique lattice of blocks. The answer is no.

In Figs. 4.8A and B we give two 3-isostatic graphs with the same lattice and associated irreducible factors (Fig. 4.8C).

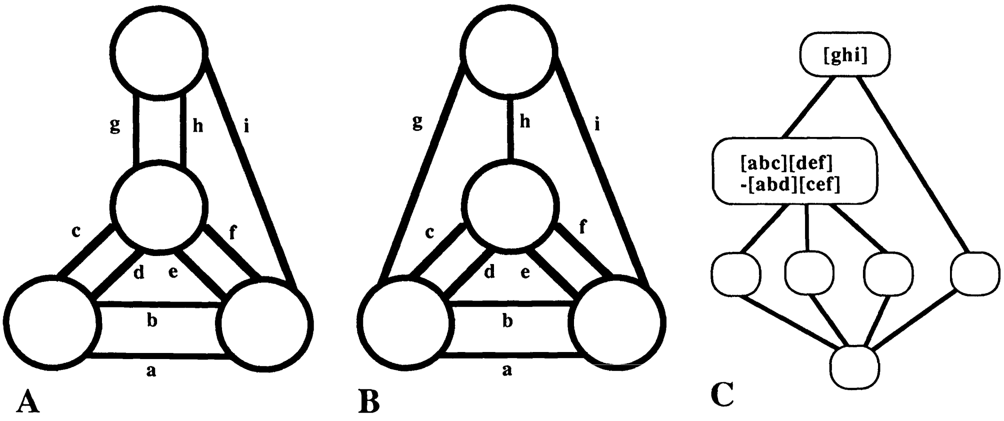


FIG. 4.8

In Fig. 4.9, we give a graph with the same pure condition, but a different lattice of blocks.

However, for each irreducible factor f, we know that there is a unique lattice for any associated k-reduced graph (a lattice with $m+1$ middle points, where f is of degree m in the brackets). We *conjecture* that there is a unique k-irreducible graph with such a factor as its pure condition.

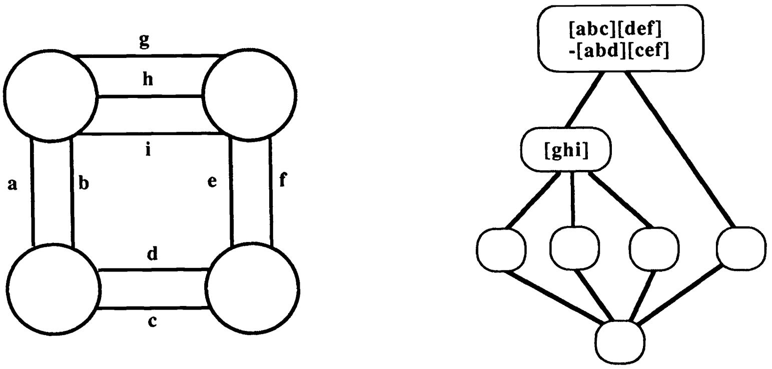


FIG. 4.9

5. Scene analysis. By our choice of topics and vocabulary, we have emphasized the role of k-frames as an abstract form of framework. Matrices with the same pattern also arise in scene analysis—the study of which plane pictures represent spatial scenes formed by distinct planes in space with designated points of contact which project to given points in the picture [19].

A basic correspondence between plane pictures, with their spatial scenes, and associated 3-frames, with their motions, has been described, in detail, in [28]. We recall three critical features of this correspondence.

(i) A picture $S(p)$ is an abstract incidence structure $S=(V, F; I)$ with vertices V, faces F, and incidences $I \subseteq VXF$, together with a mapping $p: V \to R^2(p_i=(x_i, y_i))$. Such a picture corresponds to a 3-frame with a vertex for each face, and a tree of collinear edges (with coordinates $(x_i, y_i, 1)$) spanning the faces incident to each vertex v_i with coordinates (x_i, y_i) in the picture.

(ii) A *scene* $S(q, r)$ is an assignment $q: V \to R^3(q(i)=(x_i, y_i, z_i))$ and $r: F \to R^3$, $(r(j)=(a_j, b_j, c_j))$ such that $a_jx_i+b_jy_i+z_i+c_j=0$ (the point is on the plane) if the

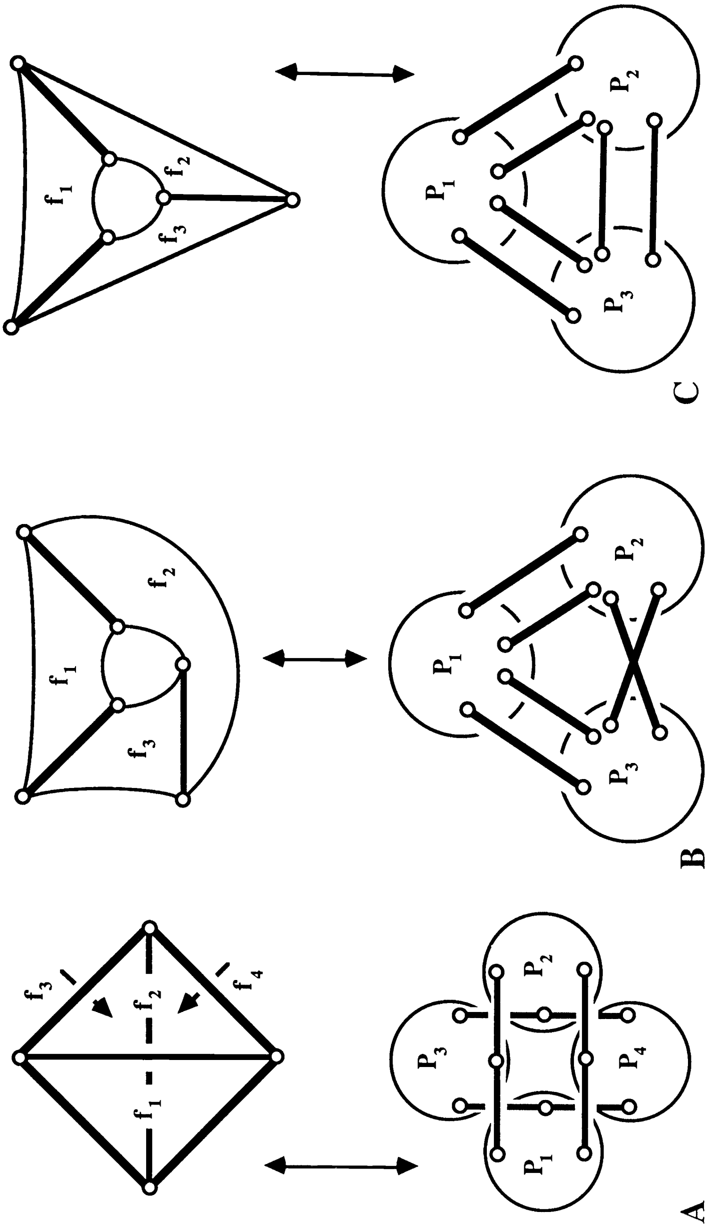


FIG. 5.1

vertex is incident with the face in I, and $q(i)$ has the same x and y coordinates as $p(i)$ (i.e. $q(i)$ projects onto $p(i)$). Such a scene over a picture corresponds to a motion of the corresponding 3-frame, where the vertex of face j has the motion $r(j)$. Flat scenes (with all faces in the same plane) correspond to trivial motions (with all bodies receiving the same center).

(iii) The desirable *sharp scenes* (with each face in a different plane) correspond to complete internal motions (with each pair of vertices in relative motion).

In Fig. 5.1 we show three pairs of corresponding pictures and frameworks. The example in Fig. 5.1A is a framework of four bodies and eight bars which corresponds to the picture of a tetrahedron. Such a picture always has a nontrivial scene, and the framework has a nontrivial motion (just by the count). The example in Fig. 5.1B is our standard framework with three bodies and six bars which is generically isostatic (Example 3.1), so the corresponding picture has only flat scenes. If the picture is drawn to represent a proper prism of three planes (Fig. 5.1C), then the corresponding framework has a complete motion.

We see from (i) that a general position picture for an incidence structure does not produce a general position 3-frame for a corresponding graph. The requirement that certain sets of edges must be collinear gives a special position to the 3-frames corresponding to pictures. When we identify the variables for edges which must be collinear, the pure condition of a 3-counted graph will specialize to the pure condition of the corresponding incidence structure. (The pure condition obtained for the incidence structure will be independent of which of the equivalent collinear trees is chosen to correspond to each vertex.) A picture $S(p)$ will have a nontrivial scene if the point $p'=(x_1, y_1, 1, \cdots, x_i, y_i, 1, \cdots)$ satisfies this pure condition $C(p')=0$.

This identification of edges causes a modification in the counting algorithm for generically correct pictures [30, Thm. 5.2]:

> An incidence structure gives a minimal flat picture if and only if $|V|+3|F|-3=|I|$, and $|V'|+3|F'|-3 \geqq |I'|$ for all proper substructures.

Of course this identification of variables also destroys the linearity of the pure condition—and may introduce many complications into the factoring of pure conditions of incidence structures. The lattice of blocks, and its associated partition of factors will remain. However most other results of § 4 (including the uniqueness of factors in a block) will be disturbed.

With these specialized matrices, Rosenberg's method for calculating conditions on k-frames can also be modified. Since no vertex should have two exiting edges in a 3-fan diagram with the same coordinates, each tree of collinear edges (corresponding to a vertex of the incidence structure) should be rooted and oriented as a whole towards this root. As a result the number of compatible 3-fans will be drastically reduced. The proof of Theorem 2.18 actually illustrated this principle, where we specialized to k trees and we were left with a single k-fan.

If we take an incidence structure whose pictures are, in general, sharp ($|V'|+3|F'|-4 \geqq I'$ for all substructures) and are maximal ($|V|+3|F|-4=|I|$) then the corresponding k-frames for these pictures will have a single motion which is computed by the methods of § 3. This method will therefore compute the planes of a general scene over such a picture. The results could be used to investigate other aspects of the picture—such as additional inequalities which follow from visual occlusion (requirements that one plane be above another at a specified point) [28, § 6].

We leave further discussion of these and other associated topics for another occasion.

REFERENCES

[1] M. BARNABEI, A. BRINI AND G. C. ROTA, *On the exterior calculus of invariant theory*, J. Algebra, 96 (1985), pp. 120-160.

[2] A. CHEUNG AND H. CRAPO, *A combinatorial perspective on algebraic geometry*, Adv. in Math., 20 (1976), pp. 388-415.

[3] R. CONNELLY, *Rigid circle and sphere packings*, I, II, Structural Topology, to appear.

[4] H. CRAPO AND W. WHITELEY, *Statics of frameworks and motions of panel structures: a projective geometric introduction*, Structural Topology, 6 (1982), pp. 43-82.

[5] P. DOUBILET, G. C. ROTA AND J. STEIN, *On the foundations of combinatorial theory:* IX *combinatorial methods in invariant theory*, Stud. Appl. Math., 53 (1974), pp. 185-216.

[6] L. HENNEBERG, *Die Graphische Statik der Starren System*, Liepzig 1911, Johnson reprint, 1968.

[7] W. V. D. HODGE AND D. PEDOE, *Methods of Algebraic Geometry Vol.* I, Cambridge University Press, London, 1946.

[8] ———, *Methods of Algebraic Geometry Vol.* II, Cambridge University Press, London, 1952.

[9] K. H. HUNT, *Kinematic Geometry of Mechanisms*, Oxford University Press, London, 1978.

[10] H. JESSOP, *A Treatise on the Line Complex*, Cambridge University Press, London, 1903, reprint, Chelsea, New York, 1969.

[11] F. KLEIN, *Elementary Mathematics from an Advanced Standpoint: Geometry*, English translation, Dover, New York, 1939.

[12] G. LAMAN, *On graphs and the rigidity of plane skeletal structures*, J. Engrg. Math., 4 (1970), pp. 331-340.

[13] H. LIPKIN AND J. DUFFY, *Analysis of industrial robots via the theory of screws*, presented at the 12th Internat. Symposium on Industrial Robots, preprint, Center for Intelligent Machines and Robotics, Univ. Florida, Gainesville, Florida, 1982.

[14] L. LOVASZ AND Y. YEMINI, *On generic rigidity in the plane*, this Journal, 3 (1982), pp. 91-98.

[15] A. RECSKI, *A network approach to the rigidity of skeletal structures* II, Discrete Appl. Math., 8 (1984), pp. 63-68.

[16] I. ROSENBERG, *Structural rigidity in the plane*, preprint C.R.M. 510, Université de Montréal, Quebec, 1975.

[17] B. ROTH AND W. WHITELEY, *Tensegrity frameworks*, Trans. Amer. Math. Soc., 265 (1981), pp. 419-446.

[18] K. SUGIHARA, *A unifying approach to descriptive geometry and mechanisms*, Discrete Appl. Math., 5 (1983), pp. 313-328.

[19] ———, *An algebraic and combinatorial approach to the analysis of line drawings of polyhedra*, Discrete Appl. Math., 9 (1984), pp. 77-104.

[20] K. SUGIMOTO, J. DUFFY AND K. H. HUNT, *Special configurations of spatial mechanisms and robot arms*, Mechanism and Machine Theory, 17 (1982), pp. 119-132.

[21] T-S. TAY, *Rigidity of multi-graphs* I, *linking rigid bodies in n-space*, J. Comb. Theory, B 36 (1984), pp. 95-112.

[22] ———, *Rigidity of multi-graphs* II, *more on linking rigid bodies*, in Graph Theory in Singapore 1983 Proceedings, Lecture Notes in Mathematics 1073, Springer-Verlag, Berlin, 1984, pp. 129-134.

[23] T-S. TAY AND W. WHITELEY, *Recent advances in the generic rigidity of frameworks*, Structural Topology, 9 (1984), pp. 31-38.

[24] N. WHITE, *The bracket of* 2-*extensors*, Congress. Numer., 40 (1983), pp. 419-428.

[25] N. WHITE AND W. WHITELEY, *The algebraic geometry of stresses in frameworks*, this Journal, 4 (1983), pp. 481-511.

[26] ———, *A class of matroids defined on graphs and hypergraphs by counting properties*, Dept. Math., Univ. Florida, Gainesville, Florida, 1984, preprint.

[27] W. WHITELEY, *Introduction to structural geometry* I: *infinitesimal motions*, Notes, Champlain Regional College, St. Lambert, Quebec, Canada, 1976.

[28] ———, *A correspondence between scene analysis and motions of frameworks*, Discrete Appl. Math., 9 (1984), pp. 269-295.

[29] ———, *The union of matroids and the rigidity of frameworks*, Champlain Regional College, St. Lambert, Quebec, Canada, 1984, preprint.

[30] ———, *A matroid on hypergraphs with applications in scene analysis and geometry*, Champlain Regional College, St. Lambert, Quebec, Canada, 1984, preprint.

[31] O. ZARISKI AND P. SAMUEL, *Commutative Algebra Vol.* II, Van Nostrand, Princeton, New Jersey, 1960.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

002

EMBEDDING GRAPHS IN BOOKS: A LAYOUT PROBLEM WITH APPLICATIONS TO VLSI DESIGN*

FAN R. K. CHUNG†, FRANK THOMSON LEIGHTON‡ AND ARNOLD L. ROSENBERG§

Abstract. We study the graph-theoretic problem of embedding a graph in a book with its vertices in a line along the spine of the book and its edges on the pages in such a way that edges residing on the same page do not cross. This problem abstracts layout problems arising in the routing of multilayer printed circuit boards and in the design of fault-tolerant processor arrays. In devising an embedding, one strives to minimize both the number of pages used and the "cutwidth" of the edges on each page. Our main results (1) present optimal embeddings of a variety of families of graphs; (2) exhibit situations where one can achieve small pagenumber only at the expense of large cutwidth; and (3) establish bounds on the minimum pagenumber of a graph based on various structural properties of the graph. Notable in the last category are proofs that (a) every n-vertex d-valent graph can be embedded using $O(dn^{1/2})$ pages, and (b) for every $d>2$ and all large n, there are n-vertex d-valent graphs whose pagenumber is at least

$$\Omega\left(\frac{n^{1/2-1/d}}{\log^2 n}\right).$$

Key words. book embeddings, arrays of processors, fault-tolerant computing

AMS(MOS) subject classifications. 05C99, 94C15, 68C25

1. Introduction.

1.1. The problem. We study here a graph embedding problem that can be viewed in a variety of ways. We start with an undirected graph G.

Formulation 1. To embed G in a book, with its vertices on the spine of the book and its edges on the pages, in such a way that edges residing on the same page do not cross.

We seek embeddings of graphs in books that use pages that are few in number and small in width. (The *width of a page* is the maximum number of edges that cross any line perpendicular to the spine of the book. The *width of a book embedding* is the maximum width of any page of the book. The *cumulative pagewidth of a book embedding* is the sum of the widths of all the pages.) The results we present are of four types:

(1) We characterize graphs that can be embedded in books having one or two pages. For instance, the one-page graphs are precisely the outerplanar graphs. (A graph is *outerplanar* if its vertices can be placed on a circle in such a way that its edges are noncrossing chords of the circle.)

(2) We find upper bounds on the number of pages required by graphs of valence (i.e., vertex-degree) at most d, and we show that these bounds are often approached

* Received by the editors December 17, 1984, and in revised form February 20, 1986. A portion of this paper was presented at the Thirteenth International IEEE Symposium on Fault-Tolerant Computing and a portion at the Fifth International Symposium on Theory and Applications of Graphs.

† Combinatorial Optimization Department, Bell Communications Research, Morristown, New Jersey 07960.

‡ Department of Mathematics and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. A portion of this research was done while this author was visiting Bell Laboratories, Murray Hill, New Jersey. The research of this author was supported in part by a National Science Foundation Presidential Young Investigator Award with matching funds from Xerox, by Defence Advanced Research Projects Agency contract N00014-80-C-0622, and by Air Force contract OSR-82-0326.

§ Department of Computer Science, Duke University, Durham, North Carolina 27706. A portion of this research was done while this author was visiting Bell Laboratories, Murray Hill, New Jersey. The research of this author was supported in part by National Science Foundation grants MCS-81-16522, MCS-83-01213 and DMC-85-04308.

by specific d-valent graphs. For example, every n-vertex $(d>2)$-valent graph can be embedded in a book with $\min(n/2, O(dn^{1/2}))$ pages (graphs of valence $d \leqq 2$ require only one page); and there exist such graphs that cannot be embedded in fewer than $\Omega(n^{1/2-1/d}/\log^2 n)$ pages. (All logarithms are to the base 2.)

(3) We find optimal or near-optimal embeddings of a variety of families of graphs, including trees, grids, X-trees, cyclic shifters, permutation networks, and complete graphs. For example, every n-vertex d-ary tree can be embedded in a book having one page, of width $\lceil d/2 \rceil \cdot \log n$.

(4) We exhibit two instances of a tradeoff between the number of pages and the widths of the pages. For example, every one-page embedding of the depth-n "ladder" graph requires width $n/2$, but there are width-2 two-page embeddings for this graph.

1.2. The origins of the problem. The problem has several origins.

Sorting with parallel stacks. Even and Itai [10] and Tarjan [24] study the problem of how to realize fixed permutations of $\{1, \cdots, n\}$ with noncommunicating stacks. Initially each number is PUSHed, in the order 1 to n, onto any one of the stacks. After all the numbers are on stacks, the stacks are POPped to form the permutation. One can view this problem graph-theoretically as follows. Say we are studying permutations of $\{1, \cdots, n\}$. Then consider the bipartite graph G_n with vertices $\{a_1, \cdots, a_n, b_1, \cdots, b_n\}$ and edges connecting each a_i to b_i. The problem of realizing the permutation π on $\{1, \cdots, n\}$ with k parallel stacks is equivalent to embedding G_n in a k-page book, with its vertices embedded in the order $a_1, \cdots, a_n, b_{\pi(1)}, \cdots, b_{\pi(n)}$.

Single-row routing. In an attempt to simplify the problem of routing multilayer printed circuit boards (PCBs), So [22] decomposed the problem in the following way. In his variant, one arranges the circuit elements in a regular grid, with wiring channels separating rows and columns of elements. One then decomposes the circuit's net lists (possibly by adding new dummy elements) so that every net connects elements in a single row or in a single column. The PCB can now be routed by routing each of its rows and each of its columns independently. The variant of this scenario that does not allow a net to run from the top of a row around to its bottom nor to change layers en route [20] corresponds directly to our embedding problem applied to small-valence graphs.

Fault-tolerant processor arrays. The DIOGENES approach to the design of fault-tolerant arrays of identical processing elements (PEs, for short) [7], [21] uses "stacks of wires" to configure around faulty PEs. In broad terms, the approach works as follows. The PEs are laid out in a (logical, if not physical) line, with some number of "bundles" of wires running above the line of PEs. One then scans along the line of PEs to determine which are faulty and which are fault-free. As each good PE is encountered, it is hooked into the bundles of wires through a network of switches, thereby connecting that PE to the fault-free PEs that have already been found and preparing it for eventual connection to those that will be found. To simplify the configuration process, each bundle is made to behave like a stack, as illustrated by the following embedding of a complete depth-d binary tree (see Fig. 1). One uses a single bundle whose wires are numbered $1, \cdots, d$. After determining which of the PEs

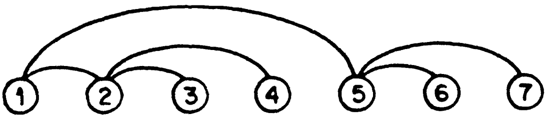


FIG. 1. *The preorder* 1*-page layout of the depth-*3 *complete binary tree.*

are good and which are faulty, one proceeds down the line of PEs from right to left. As a good PE that is to be a leaf of the tree is encountered, it is connected to line 1 in the bundle, simultaneously having lines 1 through $d-1$ "shift up," to "become" lines 2 through d; switches disconnect the left parts of the lines from the right parts so that vertex-to-vertex connectivity remains correct. The bundle has thus behaved like a stack being PUSHed. As a good PE that is to be a nonleaf of the tree is encountered, it is connected to the stack/bundle in two stages. First it is connected to lines 1 and 2 of the bundle, simultaneously having lines 3 through d "shift down" to "become" lines 1 through $d-2$; again switches ensure that proper vertex-to-vertex connectivity is maintained. The bundle here behaves like a stack being twice POPped. Second, the PE PUSHes a connection onto the stack. In this scenario, POPs amount to having a PE adopt two children that lie to its right in the line, while PUSHes amount to having the PE request to be adopted by some higher level vertex that lies to its left. The process just described lays the tree out in *preorder* and, hence, uses at most d lines.

Although not directly related to the research in this paper, the following relationship to Turing-machine graphs is also of interest.

Turing-machine graphs. One can construct a T-vertex graph that "models" a given T-step Turing machine computation, as follows. Each vertex of the graph corresponds to a step of the computation; vertices t_1 and t_2 are adjacent in the graph just if one of the machine's tape heads visits the same tape square at times t_1 and t_2, but at no intervening time. One can easily show that every k-tape Turing-machine graph is embeddable in a $2k$-page book. Hence, a characterization of graphs that are embeddable in books with a given number of pages might have applications to complexity theory. For example, a proof that such graphs have small bisection width would lead to several interesting complexity-theoretic results.

1.3. Additional formulations. Our perusal of the origins of the problem affords us additional formulations with which to hone our intuition.

Formulation 2. To place the vertices of G in a line and to assign its edges to stacks in such a way that the stacks can be used to lay out the edges.

Formulation 3. To embed the graph G so that its vertices lie on a circle and its edges are chords of the circle; to assign the chords to layers so that edges/chords on the same layer do not cross.

Formulation 3 combines the insights of [10] and [22], and yields a simple characterization of the 1-page embeddable graphs.

THEOREM 1.1 [3]. *A graph can be embedded in a one-page book if, and only if, it is outerplanar.*

Proof sketch. A graph G is outerplanar just when its vertices can be placed on a circle so that its edges become noncrossing chords of the circle.

If G is outerplanar and is laid out on a circle as above, then cutting the circle between any two vertices and opening it out to form a line yields a one-page embedding of G.

Conversely, given a one-page embedding of G, passing a line through the vertices of G in their order in the embedding and joining the ends of the line together to form a circle demonstrates G's outerplanarity. □

This characterization suggests yet another formulation.

Formulation 4. To decompose G into outerplanar graphs all of whose outerplanarity is witnessed by the same embedding of G's vertices.

1.4. Reflections from the facets. The many formulations of our problem suggest at least two variants: the first assumes that the layout of the vertices is fixed (as in sorting with parallel stacks and single-row routing); the second leaves the arrangement of the vertices as part of the problem (as in the construction of fault-tolerant processor arrays). We focus in this paper on the harder version of the problem, in which the placement of the vertices is not given.

The many facets of our problem further allow us to draw on results obtained in a variety of contexts.

The first result follows from Tarjan's analysis of the number of stacks that are required to compute a given permutation of $\{1, \cdots, n\}$. We translate the result to our graph-theoretic setting.

THEOREM 1.2 [24]. *Let the graph G have vertices* $\{a_1, \cdots, a_n, b_1, \cdots, b_n\}$ *and edges connecting each* a_i *to* b_i. *Let* π_1 *and* π_2 *be permutations of* $\{1, \cdots, n\}$. *Let the vertices of G be placed in a line in the order* $a_{\pi_1(1)}, \cdots, a_{\pi_1(n)}, b_{\pi_2(n)}, \cdots, b_{\pi_2(1)}$. *The number of pages needed to embed G given this placement of its vertices is precisely the length of the longest sequence of b-vertices whose indices are similarly ordered with their a-mates.*

The next result is immediate from the following important observation by Even and Itai [10]: The problem

> To minimize the number of pages required to embed a graph G in a book, when the ordering of G's vertices along the spine of the book is prespecified

is equivalent to the problem

> To find a minimum vertex-coloring for a *circle graph* (which is the intersection-graph for chords of a circle).

The correspondence between the two problems is best seen from Formulation 3 of the book-embedding problem. Garey et al. [13] show that the coloring problem for circle graphs is NP-complete.

THEOREM 1.3 [10], [13]. *The following problem is* NP-*complete: Given a graph G, an ordering of the vertices of G, and an integer k, decide whether or not G can be embedded in a k-page book when its vertices are placed along the spine of the book in the specified order.*

See [1] for a related result.

2. Sample embeddings and helpful principles. The problems of embedding small-valence graphs and of analyzing given embeddings are harder than they seem at first. In order to help the reader develop intuition for the remaining sections, we now present helpful strategies for obtaining bounds, and we illustrate them with sample embeddings and their analyses.

2.1. An embedding strategy. Formulation 3 of our problem suggests a strategy for embedding graphs in books, that is valuable both in finding and describing embeddings. In order to embed the graph G in a book, the strategy advocates:

1. embedding the vertices of G in a circle by finding a hamiltonian cycle in G or in some *edge-augmentation* of G (that is, a graph obtained from G by adding zero or more new edges);

2. assigning the edges of G (which are easily transformed into chords of the circle) to pages in some noncrossing manner, perhaps by coloring the vertices of the associated circle graph.

Reinforcing the intuition behind this heuristic is the fact that hamiltonian cycles add virtually no cost to an embedding: a cycle adds only 1 to the cutwidth of a layout (since one snips it), and it does not interfere with any other edges, so it does not increase the pagenumber of the embedding.

2.2. Two strategies for lower bounds. The first strategy for bounding pagenumber from below resides in the following result, which follows from Theorem 4.1 (q.v.).

THEOREM 2.1. *If the graph G is not planar, then it cannot be embedded in fewer than three pages.*

The second bounding strategy revolves around the properties of *matching graphs.* For our purposes a matching graph is a regular univalent graph (hence has an even number of vertices). If we view a matching graph as being bipartite, we can naturally associate with it a permutation π: the graph's "input" vertices are labelled $1, \cdots, n$ and are connected, respectively, to "output" vertices $\pi(1), \cdots, \pi(n)$. We shall encounter situations when analyzing a specific layout or a class of layouts of a graph G wherein we can assert that G must contain as a subgraph a matching graph G^* such that

1. the input vertices of G^* all lie to one side of its output vertices;

2. the input and output vertices of G^* are *similarly ordered,* in the sense that, if the inputs are laid out in the order $v_1, v_2, \cdots, v_n$, then the outputs appear in the order $\pi(v_1), \pi(v_2), \cdots, \pi(v_n)$.

When the existence of such a G^* can be established, we can infer that this (class of) embedding(s) of G requires n pages. The reasoning leading to this conclusion bears a strong kinship with the reasoning that Tarjan [24] and Even and Itai [10] used when studying sequences of integers that can be sorted using n stacks.

The lower bounds we obtain via matching subgraphs are among the best we derive in the paper.

2.3. Sample embeddings.

2.3.1. The pinwheel graph. The embeddings we shall be presenting in the course of our study will bear out the value of the hamiltonian-cycle embedding strategy. The following example illustrates how careful one must be to search for a *good* hamiltonian cycle.

The *depth-n pinwheel graph* $P(n)$ has $2n$ vertices

$$\{a_1, a_2, \cdots, a_n\}$$

and

$$\{b_1, b_2, \cdots, b_n\}$$

and edges connecting each pair of vertices of the form

$$\begin{aligned} &a_i - b_i, && 1 \leqq i \leqq n, \\ &a_i - b_{n-i+1}, && 1 \leqq i \leqq n, \\ &a_i - a_{i+1}, && 1 \leqq i < n, \\ &b_i - b_{i+1}, && 1 \leqq i < n. \end{aligned}$$

See Fig. 2.

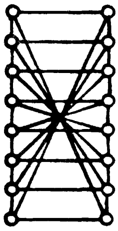


FIG. 2. *The depth-*8 *pinwheel graph* $P(8)$.

When $n > 2$, the graph $P(n)$ is not planar: $P(3) = K_{3,3}$. We shall see in § 4 (Theorem 4.1) that this nonplanarity precludes $P(n)$'s being embedded in fewer than three pages. Can one do this well? The obvious hamiltonian cycle—that goes "down the a's and up the b's"—leads to an embedding using roughly n pages, one of width proportional to n. However, if one studies the structure of pinwheels more carefully, then one discovers a hamiltonian cycle that leads to a 3-page embedding for $P(n)$, independent of n, in which the three pages have widths 2, 4, and 4, respectively.

PROPOSITION 2.2. *The graph $P(n)$ is 3-page embeddable in such a way that one page has width 2 and the other two have width 4 each.*[1]

Proof. The embedding. One obtains the desired cycle by rearranging the "butterflies" that comprise $P(n)$, as follows. We use asterisks to divide the cycle into segments that facilitate the analysis of the induced embedding. Assume for simplicity that n is even.

$$a_1 - b_1 - a_n - b_n - b_{n-1} - a_{n-1} - b_2 - a_2 - * - a_3 - b_3 - a_{n-2} - b_{n-2} - b_{n-3} - a_{n-3} - b_4 - a_4 - * -$$

$$\cdots - * - a_{n/2-1} - b_{n/2-1} - a_{n/2+2} - b_{n/2+2} - b_{n/2+1} - a_{n/2+1} - b_{n/2} - a_{n/2}.$$

Each segment of the cycle comprises two adjacent butterflies, the second recorded in reverse order of the first. Let us linearize the vertices of $P(n)$ by snipping the cycle between a_1 and $a_{n/2}$, as suggested by the way we have written the cycle.

The analysis. For each segment, we need one width-2 page to hold the butterfly edges. A second, width-4, page suffices to hold the edges that connect any single pair of adjacent butterflies. But, if this page is used for the edges that connect the ith and $(i+1)$th butterflies, it cannot also hold the edges between the $(i+1)$th and $(i+2)$th butterflies; for this next pair we need yet a third width-4 page. We need no additional pages, since the latter two can alternate joining up adjacent butterflies. Thus the cycle we have presented leads to a layout with the claimed efficiency. □

2.3.2. The sum of triangles graph. The next graph we look at is interesting because of the techniques that are needed to analyze and bound the efficiency of its embeddings. In particular, it will afford our first use of matching subgraphs to obtain a lower bound on pagenumber.

The *depth-n sum of triangles graph* $T(n)$ has vertices

$$\{a_i, b_i, c_i: 1 \leqq i \leqq n\}$$

and edges connecting each triple a_i, b_i, c_i into a triangle.

THEOREM 2.3. *The graph $T(n)$ is 1-page embeddable, with width 2. However, if one insists that $T(n)$ be laid out "by columns", so that the vertices $\{a_i\}$ are all contiguous, and so are the vertices $\{b_i\}$ and the vertices $\{c_i\}$, then $T(n)$ is $3n^{1/3}$-page embeddable, and this is optimal, within a factor of* 3.

Proof. The unrestricted layout of $T(n)$ being obvious (triangle by triangle), we restrict attention to layouts of $T(n)$ that keep all the a-vertices, all the b-vertices, and all the c-vertices contiguous, so we can refer with no ambiguity to the *a-block* of vertices, the *b-block*, and the *c-block*. We shall henceforth assume such a layout without further explicit mention. We shall also assume, for simplicity, that n is a perfect cube.

[1] One of the referees has found a 3-page embedding of $P(n)$ with pagewidths 4, 3, and 1, respectively.

The upper bound. Begin with each of the blocks of vertices in order: the a-block lies in the order

$$a_1, a_2, \cdots, a_n,$$

the b-block lies in the order

$$b_1, b_2, \cdots, b_n,$$

and the c-block lies in the order

$$c_1, c_2, \cdots, c_n.$$

Partition each of these blocks into $n^{1/3}$ *segments*, each segment being further subdivided into $n^{1/3}$ *runs* of $n^{1/3}$ vertices each. Each block thus has the form:

$$(R_1 \cdots R_d)(R_{d+1} \cdots R_e) \cdots (R_{f+1} \cdots R_g),$$

where runs are grouped by parentheses into segments. To this point, corresponding runs in corresponding segments are similarly ordered in each block.

Now begin rearranging the vertices within blocks as follows. Assume without loss of generality that the a-block lies to the left of the b-block, which lies to the left of the c-block.

(a) Leave the a-block as is.

(b) Rearrange the b-block by reversing the order of its segments, and reversing the order of the runs within each segment (but keeping vertices within runs in order, as before). The block will now look like:

$$(R_g \cdots R_{f+1}) \cdots (R_e \cdots R_{d+1})(R_d \cdots R_1).$$

(c) Rearrange the c-block by reversing the order of the runs in each segment and reversing the order of vertices within each run (but keeping the original order of the segments). If we let $\bar{R}$ denote the run obtained by reversing the vertices of the run R, then the block will now look like:

$$(\bar{R}_d \cdots \bar{R}_1)(\bar{R}_e \cdots \bar{R}_{d+1}) \cdots (\bar{R}_g \cdots \bar{R}_{f+1}).$$

Now let us add in the edges of $T(n)$ and keep track of how many pages we can get by with. When we add the edges that connect the a-block to the b-block, we note that a single page will accommodate one edge from each a-run to its corresponding b-run; since each a-run emits $n^{1/3}$ edges to the b-block, we need only this many pages to realize the a-to-b edges. When we add the edges that connect the b-block to the c-block, we note that a single page will accommodate the edges from one b-run per segment to its corresponding c-run; since there are $n^{1/3}$ runs per segment, we need only this many pages to realize the b-to-c edges. When we add the edges that connect the a-block to the c-block, we note that a single page will accommodate all the edges from one a-segment to its corresponding c-segment. Since there are only $n^{1/3}$ segments per block, we need only this many pages to implement the a-to-c edges. We have thus used $3n^{1/3}$ pages to implement all of $T(n)$'s edges.

The lower bound. Without loss of generality, say that we have $T(n)$ laid out in an a-block, a b-block, and a c-block, in that order. If we concentrate on any pair of blocks, we have a subgraph of $T(n)$ that is a matching graph whose "inputs" and "outputs" are laid out disjointly. Using the obvious correspondence between similarly (resp., oppositely) ordered inputs and outputs on the one hand, and increasing (resp., decreasing) subsequences of an integer sequence on the other hand, we note the following variant of a well-known result of Erdos and Szekeres [9].

LEMMA 2.4 [9]. *Let A and B be orderings of the integers $\{1, 2, \cdots, n\}$. If sequences A and B share no similarly ordered subsequence of length greater than k, then they share an oppositely ordered subsequence of length at least n/k.*

Now assume for contradiction that our layout of $T(n)$ requires fewer than $n^{1/3}$ pages. As we have noted in § 2.2, this implies that the a-block and the b-block share no similarly ordered subsequence of vertices of length as great as $n^{1/3}$. By Lemma 2.4, therefore, these blocks must share an oppositely ordered subsequence of length greater than $n^{2/3}$. Look now at the length-$n^{2/3}$ subsequence of the c-block that corresponds to the oppositely ordered subsequence of the a-block and the b-block. By Lemma 2.4, this subsequence of the c-block must share with the corresponding subsequence of the a-block either a similarly ordered subsequence of length

$$(n^{2/3})^{1/2} = n^{1/3}$$

or an oppositely ordered subsequence of the same length. In the former case, the edges between the a-block and the c-block cannot be realized with fewer than $n^{1/3}$ pages; in the latter case, the edges between the b-block and the c-block require this many pages. This contradicts our assumption that fewer than $n^{1/3}$ pages suffices to realize the layout of $T(n)$. □

3. Specific efficient layouts. Our attention to this point has been on establishing general analysis techniques and bounds. We now turn to the task of finding efficient layouts of a number of familiar graph families. We shall find in § 4 that these families have much more modest pagenumber demands than random graphs.

3.1. Trees. In § 1.2 we presented an embedding of the complete binary tree that turns out to be optimal in both pagenumber (one) and pagewidth ($\log n$). (Optimality of width follows from [5].) It is not hard to show that all trees enjoy embeddings that are approximately as efficient as those of complete trees.

PROPOSITION 3.1. *Every n-vertex d-ary tree can be embedded in one page of width at most*

$$\min\left(n-1, \left\lceil \frac{d}{2} \right\rceil \cdot \left\lceil \frac{\log n}{\log 3/2} \right\rceil\right).$$

Proof sketch. Let G be a graph. One *adds a fringe* to a vertex v of G by appending to v a line of (possibly 0) vertices:

$$v - v_1 - v_2 - \cdots - v_r, \qquad r \geqq 0.$$

A *fringing of G* is a graph obtained by adding a fringe to each vertex of G.

Concentrate on a single vertex v of G. Say that when G is laid out, v is flanked by vertices u and w. Let v have two fringes, $v_1, \cdots, v_r$ and $v'_1, \cdots, v'_s$ (one or both of which may be empty). Lay the fringes out either in the indicated order between v and w or in reverse order between u and v. To choose the side of v: place the first fringe on that side of v where the fewest edges of G cross or meet v (as in the conventional definition of cutwidth); place the second fringe using the same criterion in the now-augmented embedding. This strategy increases the cumulative width of the embedding by at most 1, while leaving the number of pages (one) unchanged.

An easy induction verifies that any d-ary tree T can be "built" by levels, by starting with a single vertex and "double"-fringing the graph at most

$$\left\lceil \frac{d}{2} \right\rceil \cdot \left\lceil \frac{\log |T|}{\log 3/2} \right\rceil$$

times. Our bounds on pagewidth follow from this method of constructing the tree and from the fact that the tree has at most $n-1$ edges. Details are left to the reader. □

Proposition 3.1 seeks to optimize the worst-case tree embedding. Dolev and Trickey [8] present an algorithm for finding a width-optimal one-page embedding for an individual tree.

3.2. Square grids. Square grids are planar and subhamiltonian, hence 2-page embeddable. (We verify this claim in Theorem 4.1.) The augmented hamiltonian cycle formed by row-by-row alternated east-to-west and west-to-east sweeps, as indicated in Fig. 3(a), leads to the 2-page embedding shown in Fig. 3(b). This embedding is optimal both in number of pages—the grid is not outerplanar—and in the cumulative width of the pages—the $n \times n$ grid has minimum bisection width n.

PROPOSITION 3.2. *The $n \times n$ square grid admits a 2-page embedding, each page of width n. This embedding is optimal in pagenumber and is within a factor of 2 of optimal in pagewidth.*

3.3. *X*-Trees. The *depth-d X-tree* $X(d)$ is the edge augmentation of the depth-d complete binary tree that adds edges going across each level of the tree in left-to-right order (see Fig. 4(a)).

X-trees are planar and subhamiltonian, hence admit 2-page embeddings. While it is easy to find a 2-page embedding for $X(d)$—the cycle that runs across levels in alternating orders yields one such—it is difficult to find one that has width $o(n)$ (where $n = 2^d - 1$ is the number of vertices in $X(d)$), despite the fact that $X(d)$ has a bisector of size d. However, the edge-augmentation of the X-tree depicted in Fig. 4(a), with

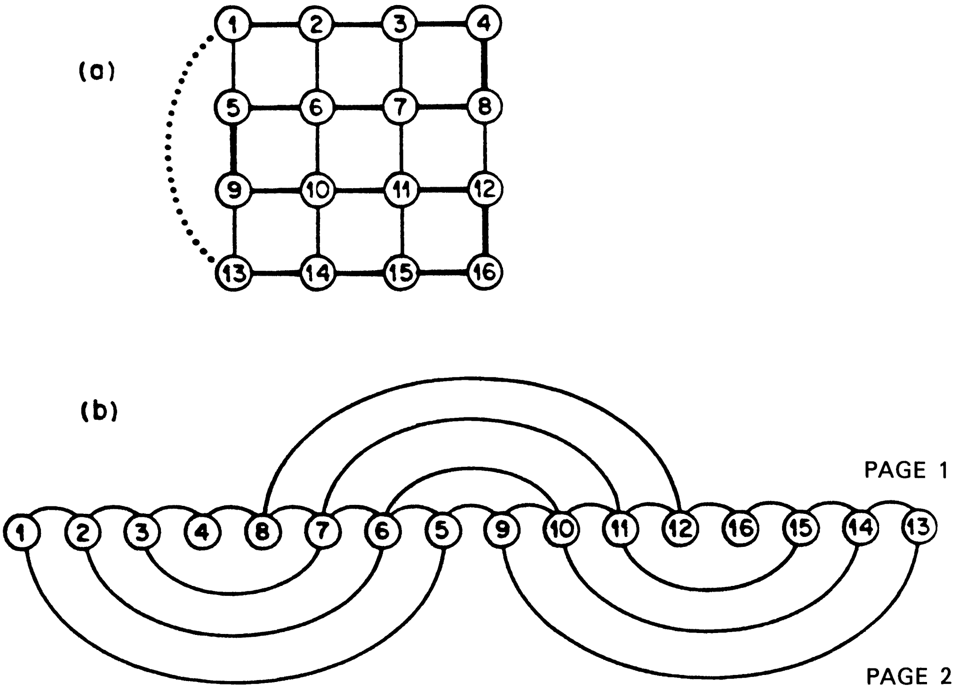


FIG. 3. (a) *The* 4×4 *grid and its efficient hamiltonian cycle.* (*In all the figures, the edges added to create an efficient cycle are shown as dotted lines; the graph edges comprising the cycles are thickened.*) (b) *The* 2-*page layout of the grid induced by the cycle.*

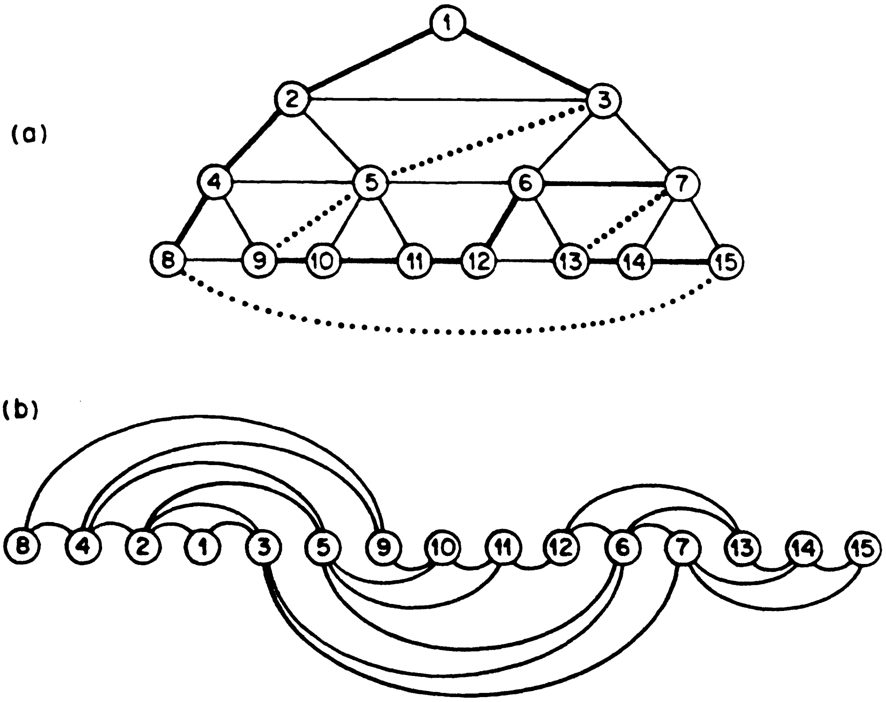


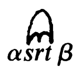
FIG. 4. (a) *An edge-augmentation of the depth-4 X-tree and an efficient hamiltonian cycle.* (b) *The layout of the X-tree induced by the cycle of* (a).

the indicated hamiltonian cycle, leads to the width-$O(d)$ 2-page embedding of $X(d)$ depicted in Fig. 4(b).

PROPOSITION 3.3. *The depth-d X-tree admits a* 2*-page embedding, with one page of width* 2*d and one of width* 3*d. This embedding is optimal in pagenumber and is within a factor of* 5 *of optimal in cumulative pagewidth.*

Proof. Optimality in number of pages is immediate since $X(d)$ is not outerplanar for $d \geqq 3$. The (near-) optimality of the claimed cutwidth follows from the proof in [17] that $X(d)$ has no bisector of size less than d, coupled with the demonstration that this implies a similar bound on cutwidth.

It remains only to verify that the widths of the pages in the prescribed embedding do indeed satisfy the claimed bounds. The verification proceeds by induction, but requires some detail about the layout of $X(d)$. Say that we have a 2-page embedding of $X(d-1)$ with the claimed pagewidths and the following form. We depict the embedding schematically by its linearization of $X(d)$'s vertices, together with a few relevant edges. For simplicity we draw page 1 above the line of vertices and page 2 below the line.



LAYOUT 1

Here r, s, t are, respectively, the root of $X(d-1)$ and its left and right sons; α and β are the strings comprising the rest of $X(d-1)$'s vertices. Assume for induction that in Layout 1:

(1) the left spine vertices (which are the leftmost vertices at each level) of $X(d-1)$ appear consecutively in leaf-to-root order in α;

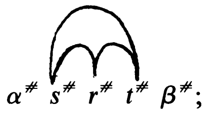
(2) the right spine vertices (which are the rightmost vertices at each level) appear, not necessarily consecutively, in root-to-leaf order in β;

(3) the vertices r, s, t and all of the left and right spine vertices are ***exposed*** on page 2, in the sense that no edge of $X(d-1)$ passes totally over them (i.e., under them in the picture);

(4) the width of page 1 is at most $2d-2$;

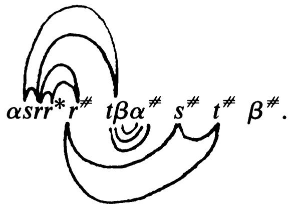
(5) the width of page 2 is 0 below the left spine vertices, and is less than $3k-3$ to the right of the level-$(d-k-1)$ spine vertices.

Now take a second copy of Layout 1:



LAYOUT 2

The prescribed layout of $X(d)$—whose set of vertices is just the union of the sets of vertices of its two depth-$(d-1)$ sub-X-trees, in addition to r^*, its root vertex—is obtained from the indicated layouts as follows:


$\alpha srr^* r^{\#}\ t\beta\alpha^{\#}\ s^{\#}\ t^{\#}\ \beta^{\#}$.

LAYOUT 3

A careful analysis of the composite layout extends the induction: Conditions (1), (2) are immediate since the left (resp., right) spine of $X(d)$ is contained in the string αsr (resp., the string $r^{\#}\ t^{\#}\ \beta^{\#}$), whose order is inherited from Layout 1 (resp., from Layout 2). Condition (3) is clear from the depiction of Layout 3: no edges are placed in the forbidden regions. Conditions (4), (5) are verified by simple counting.

Analysis of small X-trees establishes the base of the induction, thereby completing the proof. □

3.4. Benes permutation networks and their relatives. We now consider families of graphs whose structure is materially more complicated than the ones we have considered so far. These families are all very similar in structure and arise in a variety of contexts. They include the *FFT networks* whose structure represents the computational dependencies in the Fast Fourier Transform algorithm, *Banyan networks* whose structure approximates that of the Boolean n-cube while retaining bounded vertex-degrees, and the Benes rearrangeable permutation network [2], which is shown in Fig. 5(a). We concentrate on the Benes network, since it is a supergraph of the others, hence the hardest of the group to embed efficiently.

Let n be a power of 2. The *n-input Benes network* $B(n)$ is the graph defined inductively as follows.

1. $B(2)$ is the complete bipartite graph $K_{2,2}$ on two *input* vertices $i_{1,1}$ and $i_{1,2}$ and two *output* vertices $o_{1,1}$ and $o_{1,2}$.

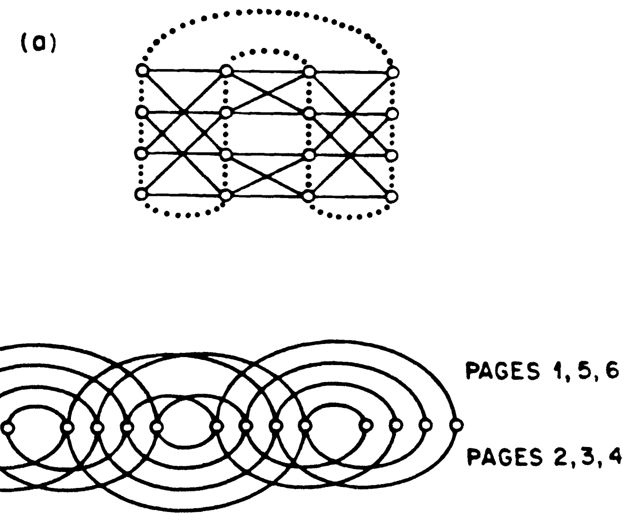


FIG. 5. (a) *The 4-input Benes network.* (b) *A 6-page layout of two levels of the network.*

2. $B(n)$ is obtained by taking two copies of $B(n/2)$ as well as n new input vertices, $i_{n,1}, i_{n,2}, \cdots, i_{n,n}$ and n new output vertices, $o_{n,1}, o_{n,2}, \cdots, o_{n,n}$. For each $1 \leqq k \leqq n$, one adds edges that create one copy of $K_{2,2}$ with "inputs" $i_{n,k}$ and $i_{n,k+n/2}$ and "outputs" $i_{n/2,k}$ and $i'_{n/2,k}$ (the primed vertices coming from the second of the two copies of $B(n/2)$) and one copy of $K_{2,2}$ with "inputs" $o_{n/2,k}$ and $o'_{n/2,k}$ and "outputs" $o_{n,k}$ and $o_{n,k+n/2}$ (again, primed vertices come from the second copy of $B(n/2)$).

Benes networks and their relatives are nonplanar, so they require at least three pages. Games [12] has recently discovered an elegant embedding that achieves this pagenumber. In order to illustrate a strategy that is often useful for finding good book embeddings, we describe now a simple 6-page embedding, which is built upon the hamiltonian cycle that alternates running up and down the "columns" of inputs and outputs of $B(n)$; see Fig. 5(b). In this embedding, one uses three pages to realize the "butterflies" that connect each "column" of vertices to the next "column." The fact that the embedding uses only a bounded number of pages is due to its reusing pages as it proceeds down the columns of $B(n)$. This strategy of reusing independent pages is a central feature of efficient embeddings (cf. [6], [15], [29]). It is somewhat surprising that any graph capable of "computing" all permutations can be realized with any bounded number, let alone 3, of pages.

PROPOSITION 3.4 [12]. *The Benes network $B(n)$ admits a 3-page embedding, with each page having width n. This embedding is optimal in pagenumber and within a factor of 3 of optimal in pagewidth.*

3.5. The Boolean *n*-cube. Our next family of graphs also has a rich interconnection structure which follows the communication structure of a broad class of algorithms. This family has been proposed as a desirable network architecture for a highly parallel computer; indeed, many of the other networks discussed in the literature—the shuffle-exchange, the banyan, and the cube-connected-cycles, for example—arose as bounded-valence stand-ins for our next graph. The *Boolean n-cube* $C(n)$ has as vertices the set of all binary strings of length n. The edges of $C(n)$ connect string-vertices x and y just when x and y are unit Hamming distance apart, i.e., when there exist binary strings

α, β, of collective length $n-1$, such that

$$\{x, y\} = \{\alpha 0\beta, \alpha 1\beta\}.$$

Thus $C(n)$ has 2^n vertices and $n2^{n-1}$ edges. Since $C(n)$ is hard to visualize for $n>3$, its efficient embedding is more easily described inductively in string-oriented terms, rather than via a hamiltonian cycle.

PROPOSITION 3.5. *The graph $C(n)$ $(n \geqq 2)$ admits an $(n-1)$-page embedding, with one page of width 2^i for each $1 \leqq i \leqq n-1$. This embedding is within a factor of 2 of optimal in both pagenumber and cumulative pagewidth.*

Proof. The lower bound on pagenumber is immediate from the facts that

(a) the pagenumber of $C(n)$ is at least as big as the minimum number of outerplanar graphs into which $C(n)$ can be decomposed (Theorem 1.1);

(b) an N-vertex outerplanar graph can have at most N "noncircle" edges [23];

(c) $C(n)$ has $n2^{n-1} = (1/2)N \cdot \log N$ edges.

The lower bound on cumulative pagewidth follows from the easily derived fact that $C(n)$ has minimum bisection width 2^{n-1}.

The upper bound is seen easily by describing inductively the linearization of the vertices of $C(n)$.

* The vertices of $C(2)$ are laid out as follows:

$$00 \quad 01 \quad 11 \quad 10$$

hence $C(2)$ is embeddable in one width-2 page.

* Assume that $C(n)$ is realized with $n-1$ pages of widths $2, 4, \cdots, 2^{n-1}$, via the linearization

$$\beta_1\beta_2 \cdots \beta_N$$

where each β_i is a distinct length-n binary word. Then the following layout for $C(n+1)$:

$$0\beta_1 0\beta_2 \cdots 0\beta_N 1\beta_N \cdots 1\beta_2 1\beta_1$$

is realizable with just one more page, of width N. This extends the induction and completes the proof. □

3.6. The complete graph K_n. Finally, we analyze the complete graph on n vertices, K_n, in which every pair of vertices is adjacent. To simplify our analysis, without losing any of the germane ideas, let us assume that n is even.

PROPOSITION 3.6. *The complete graph K_n is embeddable in $n/2$ pages, each of width at most n. This embedding is optimal in pagenumber and in cumulative pagewidth.*

Proof. We establish the claims in reverse order.

Optimality in cumulative pagewidth is immediate since, by symmetry, all layouts of K_n have the same cutwidth.

Optimality in number of pages is deducible from our principle about matching subgraphs. Lay the vertices of K_n out on a line; call the vertices $0, 1, \cdots, n-1$ in left-to-right order. Note that K_n contains as a subgraph the matching graph M_n whose input vertices are $0, 1, \cdots, (n/2)-1$, and whose output vertices are given by: $\pi(v) = v + n/2$ for $0 \leqq v < n/2$. Since the inputs and outputs of M_n are similarly ordered in this embedding, this embedding requires $n/2$ pages. Since all embeddings of K_n are isomorphic, the bound on pagenumber follows.

To see the upper bounds, consider the following way to lay out K_n. Place the vertices $0, 1, \cdots, n-1$ evenly spaced on a circle. For each vertex v, $0 \leqq v < n/2$, draw the line-graph L_v as indicated in the following illustration, in which all arithmetic is modulo n and in which double dashes denote edges of the line:

$$v = v+1 = v-1 = v+2 = v-2 = \cdots = v+(n/2)-1 = v-(n/2)+1 = v+(n/2).$$

It is not hard to verify the following facts.

(1) Each such line is composed of noncrossing chords of the circle; hence, by Theorem 1.1, each is embeddable on a single page.

(2) Every edge of K_n appears in precisely one line: to verify this, note that each vertex w is an endpoint of (hence, has valence 1 in) precisely one line, namely, $L_{w \bmod n/2}$ and has valence 2 in all other lines, so that in all, $n-1$ edges leave w; moreover, no two lines share an edge since, in the circle picture, all the lines emanating from vertex w have different slopes.

These two facts establish that, if one snips the circle between any two vertices, thereby laying K_n out in a line, and if one colors the edges of K_n according to which line L_v they lie in, one obtains an embedding of K_n in an $n/2$-page book. By the symmetry of K_n, this embedding has optimal cumulative pagewidth. □

3.7. The mesh of cliques. The $n \times n$ *mesh of cliques* $M(n)$ is the graph whose vertex-set is $\{1,2,\cdots,n\}\times\{1,2,\cdots,n\}$ and whose edges connect each row $\{i\}\times\{1,2,\cdots,n\}$ into an n-vertex clique and each column $\{1,2,\cdots,n\}\times\{i\}$ into an n-vertex clique. While we do not know how efficiently $M(n)$ can be embedded in a book in general, we can show that any embedding that places $M(n)$'s vertices along the spine row by row must use $n^{4/3}$ pages. The proof follows the inspiration of Theorem 2.3; details are left to the reader. Any nontrivial bound (particularly a lower bound) on the pagenumber of $M(n)$ would be interesting.

As a closing note to this section, Muder [18] and West [30] have a number of nontrivial bounds on the pagenumber of complete bipartite graphs $K_{n,n}$, that improve our results in [7].

4. Graph structure and pagenumber. In this section, we look at certain structural features of a graph, that are related to the number of pages required to embed the graph in a book. We find certain unexpected effects as well as the absence of certain expected ones.

4.1. Planarity. Theorem 1.1 indicates that the outerplanarity of a graph has a material effect on its pagenumber. It is easy to show that planarity has a not-dissimilar effect, but only when it is accompanied by a second structural property.

THEOREM 4.1 [3]. *The graph G admits a 2-page embedding if, and only if, it is subhamiltonian, i.e., a subgraph of a planar hamiltonian graph.*

Proof sketch. A graph is subhamiltonian just if it is embeddable in the plane so that (1) its vertices lie on a circle; (2) each of its edges lies either totally within the circle or totally without it; and (3) no edges cross in the layout.

Given such a "circular" embedding of a subhamiltonian graph G, cutting the circle between any two of G's vertices yields a planar embedding of G in a line, with each edge lying either totally above the line (i.e., on page 1) or totally below it (i.e., on page 2).

Conversely, given a 2-page embedding of the graph G, we view this embedding as placing G in a line with each edge lying totally above the line (page 1) or totally below it (page 2), and with no edges crossing. Pasting together the ends of the line containing G's vertices yields a "circular" embedding of G that witnesses G's subhamiltonian planarity. □

In the several years since the appearance of [3], the question of how many pages an arbitrary planar graph requires has attracted considerable attention. Buss and Shor [6] were the first to demonstrate that planar graphs can be embedded in a bounded number of pages; their elegant layout technique embeds an arbitrary planar graph in 9 pages. Heath [15], [16] used a quite different technique that improves this bound to 7 pages. Yannakakis [29] has recently settled the issue by proving coincident upper and lower bounds of 4 pages.

THEOREM 4.2 [29]. *Every planar graph admits a* 4*-page embedding. Moreover, there exist planar graphs requiring* 4 *pages.*

Returning to the consequences of Theorem 4.1, we observe that every series-parallel graph is 2-page embeddable. The class of *series-parallel graphs* is defined inductively as follows.

1. The 2-vertex graph with one source vertex s adjacent to one target vertex t is a series-parallel graph.

2. If G is a series-parallel graph with source vertex s and target vertex t and if G' is a series-parallel graph with source vertex s' and target vertex t', then the graph G'' obtained by "identifying" vertices t and s' is a series-parallel graph with source vertex s and target vertex t'. (This is an example of "series composition.")

3. If $G_1, \cdots, G_n$ are series parallel graphs with source vertices $s_1, \cdots, s_n$ and target vertices $t_1, \cdots, t_n$, respectively, then the graph G^* obtained by: taking a new source vertex s and adding edges between s and each of the s_i; and taking a new target vertex t and adding edges between t and each of the t_i is a series-parallel graph with source vertex s and target vertex t. (This is an example of "parallel composition.")

A graph is series-parallel just when its being so follows from provisos 1-3.

PROPOSITION 4.3. *Every series-parallel graph is* 2*-page embeddable.*

Proof. It is clear that every series-parallel graph is planar. By Theorem 4.1, then, we need only show that each such graph is subhamiltonian. This is easily proved by induction on the number of vertices in the graph, using the following inductive hypothesis.

> Given a series-parallel graph G with source vertex s and target vertex t, there is a planar edge-augmentation of G that has a hamiltonian path starting at s and ending at t.

The indicated path can then be completed to a cycle by an edge from t to s, without endangering planarity, thus establishing that the graph is subhamiltonian.

We sketch the easy induction. (1) Trivially, the unique 2-vertex series-parallel graph satisfies the claim. (2) If the graphs G and G' with source vertices s and s' and target vertices t and t' each satisfies the claim, then so also does their series composition: the desired hamiltonian path goes from s through G to t, which is identified with s', and thence through G' to t'. (3) If the graphs $G_1, \cdots, G_n$ are series-parallel, with source vertices $s_1, \cdots, s_n$ and target vertices $t_1, \cdots, t_n$, then the parallel composition of the graphs satisfies the claim: the desired hamiltonian path goes from s to s_1, thence through G_1 to t_1, to s_2, thence through G_2 to $t_2, \cdots$, from t_{n-1} to s_n, thence through G_n to t_n, and finally to t. Details are left to the reader. □

The final corollary of Theorem 4.1 is a direct consequence of Wigderson's result that the problem of deciding whether or not a maximal planar graph is hamiltonian is NP-complete [28].

COROLLARY 4.4. *The problem of deciding* 2*-page embeddability is* NP*-complete.*

4.2. Bisection width. The next structural property we consider measures the ease of recursively cutting a graph into two equal size subgraphs. We find that this measure yields a nontrivial upper bound on pagenumber but does not provide any nontrivial lower bound.

For our purposes, the simplest measure of the ease of bisecting a graph resides in the Bhatt-Leighton [4] notion of bifurcator: The graph G has a *ρ-bifurcator of size B* (B an integer and $\rho > 1$) either if G has fewer than B edges or if G admits a *decomposition tree* with the following property. The root of the tree (which is the sole vertex at level 0 of the tree) is the graph G. Each graph H at level $k \geqq 0$ of the tree that has more than one vertex gives rise to two disjoint graphs at level $k+1$, having the following properties: (a) each graph contains at least one vertex; (b) their union is H; and (c) each is connected to the other by no more than $B\rho^{-k}$ edges.

THEOREM 4.5. *If the graph G has a ρ-bifurcator of size B, then it is embeddable in $(\rho/(\rho-1))B$ pages.*

Proof. Let G have a ρ-bifurcator of size B. One begins the process of embedding G in a book by forming G's decomposition tree. One now lays G's vertices in a line (which will be the spine of the book) in the same order in which they appear as leaves of the decomposition tree. One assigns edges to pages as follows. At each level k of the tree, one creates $B\rho^{-k}$ new pages. One proceeds through all of the subgraphs of G that are split at that level, and one assigns one "cut" edge from each such subgraph to each of the new pages. No crossings can be introduced by such an assignment strategy since (a) edges that belong to the same level-k subgraph are assigned to different pages, and (b) edges that are assigned to the same page belong to disjoint intervals of vertices (because of the way vertices were laid out in the spine). It remains only to count the number of pages used in the embedding. This number is clearly bounded above by

$$\sum_{k \geqq 0} B\rho^{-k} = \left(\frac{\rho}{\rho-1}\right)B. \qquad \square$$

An immediate corollary of this result is that every small-degree n-vertex planar graph is embeddable in $O(n^{1/2})$ pages. This was the best upper bound known before the work of Buss and Shor [6], Heath [15], [16], and Yannakakis [29].

Theorem 4.5 indicates that the size of a graph's bifurcator places a nontrivial upper bound on the number of pages it requires. For the most part, this does not work in the other direction. By Theorem 4.1, every n-vertex 2-page embeddable graph has a $2^{1/2}$-bifurcator of size $O(n^{1/2})$, but once we get to 3-page embeddable graphs, knowledge of a graph's pagenumber no longer yields a nontrivial bound on the size of its bifurcators.

PROPOSITION 4.6. *There exist n-vertex 3-page embeddable graphs whose smallest ρ-bifurcators have size $\Omega(n/\log n)$ for all $\rho > 1$.*

Proof. Games [12] has shown that the n-input Benes network can be embedded in a 3-page book. A straightforward application of Thompson's lower bound proof technique [25] shows that every ρ-bifurcator of the $O(n \cdot \log n)$-vertex 3-page embeddable graph $B(n)$ has size $\Omega(n)$. $\square$

The bound in Proposition 4.6 has recently been strengthened by Galil, Kannan and Szemeredi [11], but it is still not known whether or not there exist n-vertex 3-page embeddable graphs whose smallest ρ-bifurcators have size $\Theta(n)$. As we mentioned in § 1.2, showing the existence of such graphs could have interesting consequences in classical complexity theory.

4.3. Valence. The final structural property we study is the valence of a graph. We find that this property affords us nontrivial upper and lower bounds on pagenumber.

These bounds are not very close for small or large valences, but they are close for moderate-valence graphs.

The graph G has valence d if no vertex of G has degree exceeding d. G is *regular* if all its vertices have the same degree.

4.3.1. An upper bound for *d*-valent graphs.

THEOREM 4.7. *Let d be any positive integer, and let ε be any positive constant. Say that G is a d-valent graph with n vertices, where*

$$n > \left(\frac{\ln((d+1)n^{1/2})}{\varepsilon}\right)^4.$$

If $d \leqq 2$, then G is 1-page embeddable. For any values of d and ε, G is $F(\varepsilon, d, n)$-page embeddable, where

$$F(\varepsilon, d, n) = \min\left[\frac{n}{2}, (1+\varepsilon)(2+2^{1/2})(d+1)n^{1/2}\right].$$

Proof. The cases $d \leqq 2$ are simple, for if $d = 1$, G is a matching graph, and if $d = 2$, G consists of disjoint paths and cycles.

We turn now to the case of arbitrary valence d. Say that we are given an n-vertex graph G of valence d. We note first that G is embeddable in $n/2$ pages, since K_n is (Proposition 3.6); hence we need look only at the second term in the expression for $F(\varepsilon, d, n)$. We shall justify this term (nonconstructively) by showing that not all embeddings of G in books can be "bad," in the sense of using too many pages.

We begin by decomposing G into at most $d+1$ matching graphs, $G_0, \cdots, G_d$, each having at most n vertices, by means of an edge-coloring algorithm (this is always possible by Vizing's Theorem [26]). Now consider all possible permutations of G's vertices (or, equivalently, all possible layouts of the vertices in the spine of a book).

Focus on an arbitrary permutation π and on its "behavior" on one of G's constituent matching graphs G_i. Consider those edges of G_i that connect a vertex in the left half of the layout with a vertex in the right half; say there are k such edges. These edges can be viewed (as we have noted earlier) as specifying a permutation on k integers. Since we have assumed nothing about the layout nor the edges, this permutation can be viewed as a random permutation on k integers. By a fundamental result of Hammersley [14, Thm. 6], the fraction of such permutations that have an increasing sequence of length exceeding $k^{1/2} + \varepsilon(n/2)^{1/2}$ is strictly less than

$$\exp\left(-2\varepsilon\left(\frac{n}{2}\right)^{1/2}\right).$$

This means (as we have noted before, by analogy with work of Tarjan [24]) that at most this small fraction of the layouts will require as many as $(1+\varepsilon)(n/2)^{1/2}$ pages to realize the edges of G_i that connect a vertex in the left half of the layout to a vertex in the right half (since $k \leqq n/2$).

> Recall that increasing (resp., decreasing) sequences in a permutation correspond to similarly ordered (resp., oppositely ordered) sequences of inputs and outputs of our matching graph. Moreover, one can show via a strengthened analogue of Lemma 2.4 that the existence of a length-p increasing sequence in a permutation implies that the permutation can be partitioned into p decreasing sequences. The residents of each of the pages in the layout are the edges corresponding to one of these decreasing sequences.

Now let us remove these edges that connect the two halves of the layout and their incident vertices. We are left with two (roughly) half-size copies of the same problem. Moreover, since we have been discussing a matching graph, the relative layout of the remaining vertices is completely independent of the layout of the vertices that were removed, so that once again, the permutations induced by the edges can be viewed as random ones, hence within the purview of Hammersley's theorem. This means that when we analyze each of the permutations specified by the edges that connect the left halves of each of the subgraphs with the right halves, we find that at most the fraction

$$\exp\left(-2\varepsilon\left(\frac{n}{4}\right)^{1/2}\right)$$

require as many as $(1+\varepsilon)(n/4)^{1/2}$ pages for their realization. We can now continue in this fashion to remove edges that have been considered, thereby reducing our concern to 2^i subproblems of size roughly $n/2^i$ each, each of which encounters "bad" layouts with probability less than

$$\exp\left(-2\varepsilon\left(\frac{n}{2^i}\right)^{1/2}\right).$$

We continue generating half-size subproblems until $n/2^i \leqq n^{1/2}$, for by that time, Proposition 3.6 assures us that every layout can be realized within $n^{1/2}$ pages (i.e., that the probability of a layout's being "bad" is 0). It is clear from the foregoing reasoning that the probability that a random layout requires more than

$$\sum_{i=1}^{(1/2)\log n} (1+\varepsilon)(n/2^i)^{1/2} \leqq (1+\varepsilon)(1+2^{1/2})n^{1/2}+n^{1/2}$$
$$< (1+\varepsilon)(2+2^{1/2})n^{1/2}$$

pages to realize one of G's component matching graphs is less than

$$\sum_{i=1}^{(1/2)\log n} 2^{i-1}\exp(-2\varepsilon(n/2^i)^{1/2}) \leqq n^{1/2}\exp(-\varepsilon n^{1/4}).$$

Since G is just the disjoint union of its component matching graphs, it follows that the probability that a random layout of G's vertices requires more than

$$(1+\varepsilon)(2+2^{1/2})(d+1)n^{1/2}$$

pages to realize all of G's component matching graphs, hence G itself, is no greater than

$$(d+1)n^{1/2}\exp(-\varepsilon n^{1/4}),$$

which is less than unity, by the assumed relationship among n, d, and ε.

We have thus shown that almost all orderings of G's vertices result in layouts using no more than $F(\varepsilon, d, n)$ pages. □

Remark. The result of Hammersley that is at the center of the preceding proof deals with the lengths of *monotonic* subsequences of permutations. We needed the result instantiated for *increasing* subsequences, for this yielded the sought bound on *pagenumber.* However, the result can also be instantiated for *decreasing* sequences, thereby giving an $O(n^{1/2})$ upper bound on *pagewidth* also. Details are left to the reader.

4.3.2. A construction for trivalent graphs. The (nonconstructive) upper bound of Theorem 4.7 holds for almost all orderings of the vertices of arbitrary d-valent graphs, but we do not have a general construction that yields a good ordering. If we restrict attention to trivalent graphs, then we do have such an explicit construction. We begin with a special case.

Let G be a trivalent graph, and let S be the set of its degree-3 vertices. We say that G is *trimmable* if G admits a matching whose removal leaves G with at most one degree-3 vertex.

LEMMA 4.8. *Every n-vertex trimmable trivalent graph can be embedded in a* $(\frac{3}{2}n^{1/2}+5)$*-page book, each page having width at most* $n^{1/2}$.

Proof. Let G be an arbitrary n-vertex trimmable trivalent graph. We shall embed G in a book via the following series of steps.

1. We remove a matching from G, plus at most one additional edge, in such a way as to be left with a bivalent subgraph of G: in fact, a set of vertex-disjoint cycles and paths that include all of G's vertices. This is possible since G is trimmable. Let us refer to the removed matching edges as *matched* edges.

2. We (tentatively) lay G out in a line, cycle/path by cycle/path. Then we reinsert the removed edges.

3. We partition the linearized version of G into $n^{1/2}$ contiguous blocks of $n^{1/2}$ vertices each, from left to right. (Assume for simplicity that n is a perfect square.)

4. Our next task is to rearrange our tentative layout so as to achieve the claimed pagenumber. Note that every block (save possibly one) has at most $n^{1/2}+4$ edges leaving it to any other block: at most $n^{1/2}$ matching edges and at most 4 *emerging* edges that go from the cycles/paths of this block to neighboring blocks. The one possible exceptional block is the one that had one additional edge removed with the matching; it could have that additional edge leaving it, too.

We rearrange the vertices in each block, from left to right, in the following way. For the first block, we sort the vertices in decreasing order of the block numbers to which their matched edges go. For each subsequent block: (a) we place those vertices whose matched edges go to leftward blocks to the left of those vertices whose matched edges go to rightward blocks; (b) we sort the leftgoing vertices in decreasing order of the block numbers to which their emerging edges go; (c) we sort the rightgoing vertices analogously; (d) within each group of leftgoing vertices that are going to the same block, we arrange the vertices in increasing order of the distance from the present block of their target vertex.

Analysis. The effect of the rearrangements in 4(a)–(d) is that now each of the $n^{1/2}$ blocks needs just one page to realize all of its rightgoing matched edges; each of these pages has width at most $n^{1/2}$. The edges that we have scrambled within each block lie totally within blocks of size $n^{1/2}$ each; hence, we need at most half this many additional pages to realize them: By Proposition 3.6, $m/2$ pages, each of width m, can realize the edges interconnecting any group of m vertices; moreover, since the blocks are mutually disjoint, we can use the same $\frac{1}{2}n^{1/2}$ pages to realize all of them. The (at most) $4n^{1/2}$ emerging edges can be realized using at most 4 new pages: Since we never move blocks, at most two of these edges connect a block to its right neighbor, and at most two connect the block to its left neighbor; hence, the only conflicts occur within a block, and 4 new pages can resolve these conflicts. (Two of the pages used with one block can be reused in its neighbor block.) Finally, at most one additional page is necessary, to realize the one non-matched edge of G that we may have had to remove at the beginning of the embedding. The result follows. □

With the help of a crucial observation by Lenny Heath [31], we can extend Lemma 4.8 into a $(\frac{3}{2}n^{1/2}+6)$-page embedding of arbitrary trivalent graphs.

LEMMA 4.9 [31]. *Every trivalent graph without cut-edges (i.e., edges whose removal disconnects the graph) is trimmable.*

Proof. If the trivalent graph G has no cut-edges, then every vertex of G has degree 2 or 3. Let us pair up the degree-2 vertices of G and add an edge between each pair.

This will augment G to a regular trivalent graph, unless G started with an odd number of bivalent vertices, in which case our pairing leaves us with one unmated degree-2 vertex, call it v. We handle this last contingency as follows. Let $u-v-w$ be a chain in the augmented G. (If G had fewer than three vertices, it would be univalent.) Replace v and the edges (u, v) and (v, w) by the single edge (u, w). At this point, in either of the contingencies, we have augmented G to a regular trivalent graph, possibly having multiple edges, but definitely having no cut-edges (since G had none). By a well-known result of Petersen [19], the augmented G has a perfect matching, i.e., a matching whose removal renders the graph regular bivalent. If we now restore G to its original state and consider the implications of Petersen's perfect matching, we verify easily that G is trimmable. □

THEOREM 4.10. *Every n-vertex trivalent graph can be embedded in a book with* $(\frac{3}{2}n^{1/2}+6)$ *pages. Each page, save possibly one, will have width at most* $2n^{1/2}$. *The cumulative pagewidth of the embedding will at worst be proportional to n, which cannot be improved in general.*

Proof. Let us be given an arbitrary n-vertex trivalent graph G. By removing all of G's cut-edges, we decompose G into subgraphs $G_1, G_2, \cdots, G_m$, each having no cut-edges. By Lemma 4.9, each G_i is trimmable; hence, by Lemma 4.8, each G_i can be embedded in a $(\frac{3}{2}n^{1/2}+5)$-page book, each page having width at most $n^{1/2}$. Thus, any embedding of G that lays the G_i out disjointly along the line has the claimed efficiency. To prove the theorem, then, we need only show how to deal with the removed cut-edges.

We begin with two easily verified but crucial observations for which we are grateful to Lenny Heath. First, we note that if we take our layout of one of the G_i and shift the vertices cyclically, we do not change the pagenumber of the layout, and we at most double its pagewidth (since our layouts really are in circles, not lines; cf. Theorem 1.1). Second, we note that if we contract each subgraph G_i to a point, leaving only the cut-edge interconnections, then the resulting contraction of G is a tree.

Our strategy is to lay G out as a tree of subgraphs, with each subgraph laid out as in Lemma 4.8, but possibly cyclically shifted.

We begin by arbitrarily picking G_1 as the first subgraph to process. We lay G_1 out as in Lemma 4.8. Say that in the layout, the vertices

$$v_{11}, v_{12}, \cdots, v_{1k_1},$$

appearing in that order, are connected to other subgraphs by cut-edges. We place those k_1 subgraphs along the line in the reverse order of the v_{1j}. When we place each subgraph, we use the layout prescribed by Lemma 4.8; but we cyclically shift the vertices in this layout so that the leftmost cut-edge-bearing vertex is the one connected to G_1. The subgraphs just placed will remain in this order, and their layouts will stay fixed, but other subgraphs may be placed between them.

Next, we process the just-placed subgraphs recursively, from left to right. (By "recursively" here we mean the following. If we have subgraphs A and B remaining to be processed, in that order, and if in the course of processing A we place a new subgraph C between A and B, then C gets processed before B.) We process subgraph G_i, $i>1$, as follows. Say that in the layout of G_i the vertices

$$v_{i1}, v_{i2}, \cdots, v_{ik_i},$$

appearing in that order, are connected to other subgraphs by cut-edges. We place those k_i subgraphs along the line in the reverse order of the v_{ij}, *immediately to the right of* G_i (hence, to the left of all other subgraphs that have previously been placed to the right of G_i). As before, when we place each subgraph, we use the layout prescribed by Lemma 4.8; but we cyclically shift the vertices in this layout so that the leftmost cut-edge-bearing vertex is the one connected to G_i. Again, the subgraphs just placed will remain in this order, and their layouts will stay fixed, but other subgraphs may be placed between them.

The reader will recognize that we have essentially laid the contracted tree version of G out in preorder. By Proposition 3.1, then, we need only one extra page to accommodate the cut-edges. Since the contracted tree has at most n edges, the extra page has cutwidth at most n.

We thus have an embedding of G with the parameters advertised in the statement of the theorem. The cumulative pagewidth of the embedding (which is at worst proportional to n) cannot be improved in general, as one can verify by observing that the cutwidth of a trivalent n-superconcentrator must be proportional to n. □

4.3.3. A lower bound for d-valent graphs. We have been unable to find lower bounds on the worst-case pagenumber of d-valent graphs that match the upper bounds of Theorem 4.7 and Theorem 4.10. We have, however, found nontrivial lower bounds, that we present now.

THEOREM 4.11. *For all valences $d > 2$, for all sufficiently large n, there are n-vertex graphs of valence d whose pagenumber is no less than*

$$(\text{const})\,\frac{n^{(1/2)-(1/d)}}{\log^2 n}.$$

Proof. Let the valence $d > 2$ of the graphs of interest be fixed. Imagine that we have a table each of whose rows is labeled with one of the $n!$ permutations of n items ($=$ layouts of n vertices), and each of whose columns is labeled with one of the n-vertex matching graphs: the table entry corresponding to row i and column j is "FEW" if layout i uses no more than p pages on matching graph j, and is "MANY" if the layout uses more than p pages. The general strategy of our proof is to demonstrate that if p is no larger than indicated in the statement of the theorem, then some d-tuple of columns encounters at least one "MANY" in every row.

In order to get the argument going, we need to know roughly how many rows/permutations/layouts contain a "FEW" for a given column. This information is derivable from the following lemmas.

LEMMA 4.12. *At most p^{2r} permutations of r integers have no increasing sequence of length $p+1$.*

Proof. We noted in Lemma 2.4 that any permutation of $\{1, 2, \cdots, r\}$ whose longest increasing subsequence is of length p can be partitioned into p decreasing subsequences. This decomposition can be used to specify the permutation uniquely via two length-r strings over the alphabet $\{1, 2, \cdots, p\}$. The first string specifies, for each position i, which decreasing sequence occupies that position. The second string assigns the integers $\{1, 2, \cdots, r\}$ to subsequences. Since there are p^{2r} pairs of length-r strings over $\{1, 2, \cdots, p\}$, the lemma follows. □

LEMMA 4.13. *Let G be an n-vertex matching graph. The number of layouts of G that use at most p pages does not exceed*

$$P(n, p) = 2^{E(n,p)}$$

where

$$E(n, p) \leq \frac{n}{2}\log n + n\cdot\log p + 2n\cdot\log\log n.$$

Proof. Let us count the number of layouts of G that require at most p pages. We employ the correspondence we have established between matching graphs and permutations (§ 2.2). Consider an arbitrary layout of G that has r edges passing between the leftmost $n/2$ vertices of G and the rightmost $n/2$ vertices; there are obviously no more than $n/2$ such edges. Let $\binom{x}{y}$ denote the binomial coefficient

$$\binom{x}{y}=\frac{x!}{y!(x-y)!}.$$

1. There are at most $\binom{n/2}{r}$ ways to choose the r edges that cross the center of the layout.

2. Each association (= edge) between element i and element j in a permutation can arise because $\pi(i)=j$ or because $\pi(j)=i$; hence there are 2^r ways of assigning left and right halves to each of the r edges.

3. There are at most $\binom{n/2-r}{(n/2-r)/2}$ ways to assign edges that do not cross the center to either the right or the left half of the layout.

4. Since the edges that cross the center can appear in any order, there are $r!$ ways of ordering the left endpoints of these edges.

5. By Lemma 4.12, no more than p^{2r} of the permutations specified by the r edges can be realized with only p pages, so there are at most p^{2r} ways of ordering the right endpoints of the edges that cross the center.

6. There are $\binom{n/2}{r}$ ways to place the (now ordered) endpoints of the r crossing edges on each side of the layout.

Aggregating all of these possibilities, recursing down to handle the two induced subgraphs of G to the left and to the right of the center of the layout, and allowing r to range over its possible values, we end up with the recurrence

$$P(n,p)\leqq \sum_{r\leqq n/2}\binom{n/2}{r}\cdot 2^r\cdot\binom{(n/2)-r}{[(n/2)-r]/2}\cdot r!\cdot p^{2r}\cdot\binom{n/2}{r}^2\cdot\left[P\left(\frac{n}{2}-r,p\right)\right]^2.$$

Our strategy will be to take the largest term T (say that it is the rth term) from this sum and show that nT, which certainly is no less than $P(n,p)$, is no greater than the claimed bound. We begin by representing r as

$$r=b\frac{n}{2},\qquad 0<b\leqq 1,$$

and by applying to T standard estimates for the binomial coefficients. We find that

$$\begin{aligned}P(n,p)&\leqq nT\\&\leqq \exp 2\left[\log n+\tfrac{3}{2}H(b)n+\frac{n}{2}+b\frac{n}{2}\log\left(b\frac{n}{2}\right)-b\frac{n}{2}\log e+bn\log p\right.\\&\qquad\left.+2E\left((1-b)\frac{n}{2},p\right)\right]\end{aligned}$$

where $\exp 2(x)=_{\text{def}}2^x$, and where $H(b)$ is the base-2 entropy function

$$H(b)=-[b\log b+(1-b)\log(1-b)].$$

Let us now assume for induction that our claimed bound

$$E(m,p)\leqq\frac{m}{2}\log m+m\log p+2m\log\log m$$

on E (hence on P) holds for all $m<n$. It then follows from the preceding inequalities, after simplification, that

$$P(n,p) \leqq \exp 2\left[\log n + H(b)n + \frac{n}{2}\log n - b\frac{n}{2}\log e + n\log p\right.$$
$$\left. + 2(1-b)n\log\log\left((1-b)\frac{n}{2}\right)\right].$$

Note that the right-hand expression can be shown to be less than

$$\exp 2\left[\frac{n}{2}\log n + n\log p + 2n\log\log n\right]$$

provided only that for all $0<b\leqq 1$,

$$H(b) + \frac{\log n}{n} \leqq 2\log\log n - 2(1-b)\log\log\left((1-b)\frac{n}{2}\right) + \frac{b}{2}\log e.$$

We establish this last inequality by verifying that, in fact,

$$\text{(1)} \qquad H(b) + \frac{\log n}{n} \leqq \frac{2}{\log n} + 2b\log\log\left(\frac{n}{2}\right) + \frac{b}{2}\log e.$$

This will suffice since

$$\begin{aligned} 2\log\log n - 2(1-b)\log\log\left((1-b)\frac{n}{2}\right) &> 2\log\log n - 2(1-b)\log\log\left(\frac{n}{2}\right) \\ &= 2\log\log n - 2\log\log\left(\frac{n}{2}\right) + 2b\log\log\left(\frac{n}{2}\right) \\ &= 2\log\log n - 2\log(\log n - 1) + 2b\log\log\left(\frac{n}{2}\right) \\ &> \frac{2}{\log n} + 2b\log\log\left(\frac{n}{2}\right). \end{aligned}$$

Now we must verify the final inequality (1) involving $H(b)$: Using the Taylor's series expansion for $\log(1-b)$, one can show that

$$H(b) \leqq b\log\frac{1}{b} + b\log e$$

for all $b\leqq 1$. Hence it suffices to verify that

$$b\log\frac{1}{b} + \frac{b}{2}\log e + \frac{\log n}{n} \leqq \frac{2}{\log n} + 2b\log\log\left(\frac{n}{2}\right).$$

This is easily accomplished by analyzing the two cases

$$b\leqq(\log n)^{-3/2} \quad\text{and}\quad b>(\log n)^{-3/2}.$$

Thus we establish the desired inequality (1) on $H(b)$ and, through it, the desired inequality on $P(n,p)$. □

Return to proof of Theorem 4.11. Consider again our large table with entries "FEW" and "MANY". The number of "FEW" entries in each (n!-item) column of the table is at most $P(n,p)$, where p is the number of pages we are prepared to use to lay out our n-vertex d-valent graphs. Clearly, we cannot lay out all such graphs unless every d-tuple of table columns contains only "FEW" entries in at least one row of the table. (The d-tuples of this last assertion arise from the fact that every union of d matching graphs forms a d-valent graph.) These "FEW" entries have a chance of existing only if

$$c^d \leqq n!\left(\frac{P(n,p)}{n!}c\right)^d,$$

where c denotes the number of n-vertex matching graphs. The left-hand quantity is the number of d-tuples of matching graphs, while the right-hand quantity is the product of the number of rows and the number of d-tuples of "FEW" entries in each row. (The latter fact follows from the observation that, by symmetry, every row has the same number of "FEW" entries.) Simplifying, then, we can accommodate all d-valent graphs in p pages only if

$$P(n,p)^d \geqq (n!)^{d-1}.$$

By Lemma 4.13, this inequality implies (after taking logarithms)

$$dn\cdot[\tfrac{1}{2}\log n+\log p+2\log\log n]\geqq(d-1)n\log n+\Theta(n).$$

The validity of this inequality finally implies the claimed lower bound on p, namely,

$$p\geqq(\text{const})\frac{n^{1/2-1/d}}{\log^2 n}. \qquad \square$$

Our upper and lower bounds are within a few logarithmic factors apart when the valence d is logarithmic in n; they are rather far apart when d is either very big or very small. We conjecture that one of the factors of log n can be removed in the lower bound, but the tighter analysis needed is likely to be quite complicated.

5. Cost tradeoffs. In this section, we point out a rather interesting anomaly that could be important in the context of our study. We describe here two families of graphs that engender pagenumber-pagewidth *tradeoffs*. Each of these families can be laid out using some number p pages—but only if the widths of the pages are allowed to grow proportionally to the size of the graph being laid out. However, if one uses just one additional page, then the widths of the pages can be kept bounded by a constant.

Both of the graph families have the following form. The *depth-k K_n-cylinder* $C(k,n)$ is the graph whose vertex-set is the union of the k sets

$$V_{i,n}=\{v_{i,1},v_{i,2},\cdots,v_{i,n}\},\qquad 1\leqq i\leqq k,$$

and whose edges (a) connect each set $V_{i,n}$ into an n-clique, and (b) connect each vertex $v_{i,j}$ to vertex $v_{i+1,j}$, $1\leqq i<k$, $1\leqq j\leqq n$.

The anomalies of interest appear in the first two parts of the next result. The third part of the result indicates the failure of the obvious generalization of the first two parts.

PROPOSITION 5.1. (1a) *Any 1-page layout of $C(k,2)$ has pagewidth at least $k/2$.* (1b) *There are 2-page layouts of $C(k,2)$ having pagewidth 2.*

(2a) *Any 2-page layout of $C(k,3)$ has pagewidth at least $k/2$.* (2b) *There are 3-page layouts of $C(k,3)$ having pagewidth 4.*

(3) *There are 3-page layouts of $C(k,4)$ having pagewidth 4.*

Proof sketch. The fact that $C(k,2)$ is outerplanar guarantees that it is 1-page embeddable. The fact that $C(k,3)$ is planar and subhamiltonian (a hamiltonian cycle can be traced by going from $v_{1,1}$ to $v_{1,2}$ to $v_{2,1}$ to $v_{2,2}$, and so on until one has reached $v_{2,n}$; at that point one goes to $v_{n,3}$, thence to $v_{n-1,3}$, and so forth, to $v_{1,3}$) guarantees that it is 2-page embeddable. Proving the lower bounds on the pagewidths of the resulting layouts proceeds by showing that at least half of the constituent n-cliques must be nested in any minimal-page layout. This is easily verified directly in the case of $C(k, 2)$: any (not necessarily contiguous) sequence of the form

$$v_{a,1} \cdots v_{b,2} \cdots v_{c,1} \cdots v_{d,2}$$

(or its reversal), where $\{a, b\} = \{1, 2\}$ precludes an embedding using just one page. (This verification is a special case of Syslo's result [23] that every biconnected outerplanar graph has a unique outerplanar embedding.) In the case of $C(k, 3)$, a direct verification is a bit more difficult; but the result follows immediately from Whitney's proof [27] that every triconnected planar graph has a unique planar embedding.

The existence of the claimed small-pagewidth layouts can be verified by the reader from the illustrative layouts depicted in Fig. 6. □

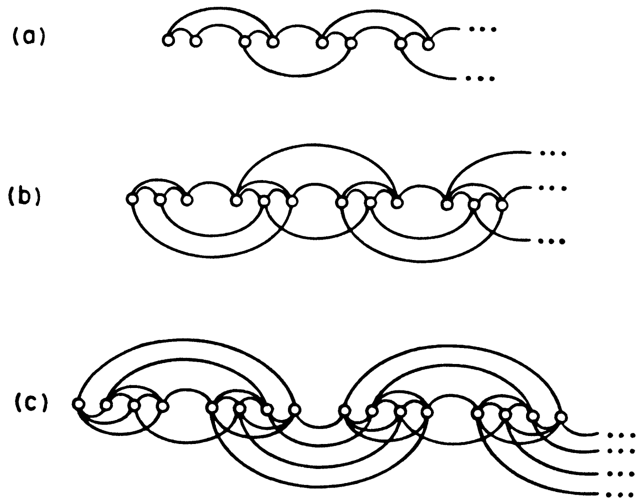


FIG. 6. *A small-width layout for* (a) $C(4,2)$, (b) $C(4,3)$, (c) $C(4,4)$.

It would be interesting to know whether or not there exist pagewidth-pagenumber tradeoffs analogous to those of Proposition 5.1 for every number of pages; i.e., can using one more page decrease pagewidth unboundedly?

Acknowledgments. It is a pleasure to thank Lenny Heath, Ravi Kannan, and Gary Miller for helpful conversations leading to several key insights.

REFERENCES

[1] P. B. ARNOLD (1982), *Complexity results for single-row routing*, Harvard Univ. Ctr. Res. Comput. Tech. Report TR-22-82.

[2] V. E. BENES (1964), *Optimal rearrangeable multistage connecting networks*, Bell Syst. Tech. J., 43, pp. 1641-1656.

[3] F. BERNHART AND P. C. KAINEN (1979), *The book thickness of a graph*, J. Combin. Theory, Ser. B, 27, pp. 320-331.

[4] S. N. BHATT AND F. T. LEIGHTON (1984), *A framework for solving* VLSI *graph layout problems*, J. Comput. Syst. Sci., 28, pp. 300-343.

[5] R. P. Brent and H. T. Kung (1980), *On the area of binary tree layouts*, Inform. Process. Lett., 11, pp. 44-46.

[6] J. Buss and P. Shor (1984), *On the pagenumber of planar graphs*, 16th Annual ACM Symp. on Theory of Computing, pp. 98-100.

[7] F. R. K. Chung, F. T. Leighton and A. L. Rosenberg (1983), DIOGENES—*A methodology for designing fault-tolerant processor arrays*, 13th Internat. Conf. on Fault-Tolerant Computing, pp. 26-32.

[8] D. Dolev and H. Trickey (1982), *Embedding a tree on a line*, IBM Report RJ-3368.

[9] P. Erdös and E. Szekeres (1935), *A combinatorial problem in geometry*, Compositio Math., 2, pp. 463-470.

[10] S. Even and A. Itai (1971), *Queues, stacks, and graphs*, in Theory of Machines and Computations, Z. Kohavi and A. Paz, eds., Academic Press, New York, pp. 71-86.

[11] Z. Galil, R. Kannan and E. Szemeredi (1986), *On nontrivial separators for k-page graphs and simulations by nondeterministic one-tape Turing machines*, 18th Annual ACM Symp. on Theory of Computing.

[12] R. A. Games (1986), *Optimal book embeddings of the FFT butterfly, Benes, and barrel shifter networks*, Algorithmica, to appear.

[13] M. R. Garey, D. S. Johnson, G. L. Miller and C. H. Papadimitriou (1980), *The complexity of coloring circular arcs and chords*, this Journal, 1, pp. 216-227.

[14] J. M. Hammersley (1972), *A few seedlings of research*, 6th Berkeley Symp. on Math. Stat. and Prob., Vol. 1, pp. 345-394.

[15] L. S. Heath (1984), *Embedding planar graphs in seven pages*, 25th IEEE Symp. on Foundations of Computer Science, pp. 74-83.

[16] ——— (1985), *Algorithms for embedding graphs in books*, Ph.D. dissertation, Univ. North Carolina, Chapel Hill, NC.

[17] J.-W. Hong and A. L. Rosenberg (1982), *Graphs that are almost binary trees*, SIAM J. Comput., 11, pp. 227-242.

[18] D. J. Muder (1985), *Book embeddings of regular complete bipartite graphs*, Typescript, The MITRE Corp.

[19] J. Petersen (1891), *Die Theorie der regulaeren Graphen*, Acta Math., 15, pp. 193-220.

[20] R. Raghavan and S. Sahni (1983), *Single row routing*, IEEE Trans. Comput., C-32, pp. 209-220.

[21] A. L. Rosenberg (1983), *The Diogenes approach to testable fault-tolerant arrays of processors*, IEEE Trans. Comput., C-32, pp. 902-910.

[22] H. C. So (1974), *Some theoretical results on the routing of multilayer printed-wiring boards*, 1974 IEEE Intl. Symp. on Circuits and Systems, pp. 296-303.

[23] M. M. Syslo (1979), *Characterizations of outerplanar graphs*, Discrete Math., 26, pp. 47-53.

[24] R. E. Tarjan (1972), *Sorting using networks of queues and stacks*, J. Assoc. Comput. Mach., 19, pp. 341-346.

[25] C. D. Thompson (1980), *A Complexity Theory for* VLSI, Ph.D. thesis, Carnegie-Mellon Univ., Pittsburgh, PA.

[26] V. G. Vizing (1964), *On an estimate of the chromatic class of a p-graph*, Diskret. Analiz., 3, pp. 25-30. (In Russian.)

[27] H. Whitney (1933), *A set of topological invariants for graphs*, Amer. J. Math., 55, pp. 221-235.

[28] A. Wigderson (1982), *The complexity of the hamiltonian circuit problem for maximal planar graphs*, Princeton Univ. EECS Dept. Report 298.

[29] M. Yannakakis (1986), *Four pages are necessary and sufficient for planar graphs*, 18th Annual ACM Symp. on Theory of Computing.

[30] D. West (1985), Personal communication.

[31] L. S. Heath (1985), Personal communication.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

003

ON THE SINGULAR "VECTORS" OF THE LYAPUNOV OPERATOR*

RALPH BYERS† AND STEPHEN NASH‡

Abstract. For a real matrix A, the separation of A^T and A is $\operatorname{sep}(A^T,-A)=\min\|A^TX+XA\|/\|X\|$, where $\|\cdot\|$ represents the Frobenius matrix norm. We discuss the conjecture that the minimizer X is symmetric. This conjecture is related to the numerical stability of methods for solving the matrix Lyapunov equation. The quotient is minimized by either a symmetric matrix or a skew-symmetric matrix and is maximized by a symmetric matrix. The conjecture is true if A is 2-by-2, if A is normal, if the minimum is zero, or if the real parts of the eigenvalues of A are of one sign. In general the conjecture is false, but counterexamples suggest that symmetric matrices are nearly optimal.

Key words. Lyapunov equation, sep, singular values, singular vectors

AMS(MOS) subject classifications. 65F35, 15A18, 15A45, 49B99, 65G99

1. Introduction.

For $A\in\mathbf{R}^{n\times n}$ define the separation $\operatorname{sep}(A^T,-A)$ [9], [10] by

$$\operatorname{sep}(A^T,-A)=\min_{X\neq 0}\frac{\|A^TX+XA\|}{\|X\|}. \tag{1}$$

Here $\|\cdot\|$ is the Frobenius matrix norm $\|Z\|^2=\operatorname{trace}(Z^HZ)$. We are concerned with the following:

CONJECTURE. The minimum in (1) is obtained by a real symmetric matrix X.

Of course, there may be nonsymmetric minimizers as well.

The sensitivity of the solutions of Lyapunov equations [7] and algebraic Riccati equations [1], [4] to perturbations in the data is governed by $\operatorname{sep}(A^T,-A)$. Related quantities govern the sensitivity of invariant subspaces [9].

In this paper we show that (1) is maximized by a symmetric matrix and is minimized by either a symmetric matrix or a skew-symmetric matrix. It is definitely minimized by a symmetric matrix if A is 2-by-2, if A is normal, if $\operatorname{sep}(A^T,-A)=0$ or if the real parts of the eigenvalues of A are of one sign. A counterexample shows that the conjecture is false.

In terms of the Lyapunov operator

$$L(X)=A^TX+XA$$

the conjecture is that the smallest singular "vector" of L is symmetric. The structure of the Lyapunov operator has been studied extensively. Some surveys on this and more general operators appear in [8] and [11].

Section 2 points out some of the conjecture's practical consequences in the design of numerical algorithms. Section 3 proves the conjecture for several special cases. Section 4 presents some counterexamples. The examples suggest that nearly optimal symmetric matrices always exist for the problem (1).

2. Applications to numerical software.

Consider the Lyapunov equation

$$L(X)=A^TX+XA=B, \tag{2}$$

where A and $B=B^T$ are known n-by-n matrices and $X=X^T\in\mathbf{R}^{n\times n}$ is unknown. If the spectra of A and $-A$ are disjoint, then there is a unique solution X. Using t-digit

* Received by the editors June 3, 1985; accepted for publication (in revised form) March 13, 1986.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

‡ Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218.

The work of this author was supported in part by National Science Foundation grant ECS-8305304.

base b arithmetic, a sound numerical algorithm for solving (2) will produce an approximate solution $\tilde{X}$ such that

$$A^T\tilde{X}+\tilde{X}A=B-R, \tag{3}$$

where the residual R is small in the sense that for some modest constant c depending on the size of the problem and the algorithm used,

$$\|R\|\leqq cb^{1-t}\|A\|\,\|\tilde{X}\|. \tag{4}$$

Such an approximate solution $\tilde{X}$ may be bounded in terms of the correct solution X of (2) using

$$A^T(X-\tilde{X})+(X-\tilde{X})A=R.$$

Thus,

$$\|X-\tilde{X}\|\leqq\frac{\|R\|}{\operatorname{sep}(A^T,-A)}. \tag{5}$$

The Bartels-Stewart algorithm [2] and the Golub-Nash-Van Loan algorithm [7] both produce approximate solutions satisfying (3) and (4), but in the presence of rounding errors, the Bartels-Stewart algorithm preserves the symmetry of X and R while the Golub-Nash-Van Loan algorithm usually does not. The latter algorithm could symmetrize the solution, but it was designed for nonsymmetric systems, and is less efficient than Bartels-Stewart in the symmetric case.

So, the Bartels-Stewart algorithm produces a better quality approximate solution. Furthermore, if (1) is not minimized by a symmetric matrix, then (5) cannot be an equality for the Bartels-Stewart algorithm. The main issue in (5) is really the difference in magnitude between the left- and right-hand sides, so this point may not be significant. (The Bartels-Stewart algorithm also requires less work and storage to solve (2) than the Golub-Nash-Van Loan algorithm. However, for the more general problem in which A^T is replaced by some other matrix the Golub-Nash-Van Loan algorithm becomes less expensive.)

An economical estimator for sep $(A^T,-A)$ described in [3] uses a heuristic similar to the LINPACK condition estimator [5] to choose an approximate minimizer of (1). The heuristically chosen minimizer can be improved by using inverse iteration on the Lyapunov operator (2) and its transpose. The approximate minimizer is symmetric and symmetry is preserved by inverse iteration. Separation estimators like [3] that use symmetric approximate minimizers may fail when (1) is not minimized by a symmetric matrix X.

3. Special cases. In this section we show that the quotient in (1) is maximized by a symmetric matrix and is minimized by either a symmetric matrix or a skew-symmetric matrix. We establish that it is minimized by a symmetric matrix if A is 2-by-2, A is normal, sep $(A^T,-A)=0$ or if the real parts of the eigenvalues of A are of one sign.

In what follows, we use the usual ordering of (possibly complex) Hermitian matrices: $B\geqq C$ if and only if $B-C$ is positive semi-definite. The notation A^H represents the Hermitian transpose of A.

The first lemma shows that the conjecture can be false only if (1) is minimized by a skew-symmetric matrix.

LEMMA 1. *Let A be a real n-by-n matrix. There exists a real matrix $Z \in \mathbf{R}^{n \times n}$ such that $Z \neq 0$,*

$$\min_{X \neq 0} \frac{\|A^T X + XA\|}{\|X\|} = \frac{\|A^T Z + ZA\|}{\|Z\|}, \tag{6}$$

and either $Z = Z^T$ or $Z = -Z^T$.

Proof. The Lyapunov operator $L(X)$ in (2) is a linear transformation on the finite-dimensional, real vector space $\mathbf{R}^{n \times n}$. The Frobenius norm is simply the vector 2-norm applied to "vectors" in $\mathbf{R}^{n \times n}$, so the minimum in (6) occurs when X is a singular "vector" of L corresponding to the smallest singular value. Let $W \in \mathbf{R}^{n \times n}$ be such a singular "vector." Since

$$\|A^T W + WA\| = \|(A^T W + WA)^T\| = \|A^T W^T + W^T A\|,$$

W^T is also a singular "vector" of L corresponding to the smallest singular value. If $W + W^T = 0$, then $Z = W$ is a skew-symmetric matrix satisfying the theorem. Otherwise, $Z = (W + W^T)$ is a symmetric singular "vector" of L corresponding to the smallest singular value. □

Notice that the set of real, symmetric matrices and the set of real, skew-symmetric matrices form invariant subspaces of the Lyapunov operator $L(X)$ (2). These subspaces are orthogonal with respect to the Frobenius inner product

$$\langle A, B \rangle = \text{trace}\,(A^T B)$$

from which the Frobenius norm arises. Since the symmetric and skew-symmetric matrices span $\mathbf{R}^{n \times n}$, the singular "vectors" of $L(X)$ may be chosen from these invariant subspaces. The interpretation in terms of singular values and the orthogonality of these two invariant subspaces makes the use of the Frobenius norm in (1) more natural than other matrix norms.

Much of what follows depends on the following corollary to the proof of Lemma 1.

COROLLARY 2. *If* (1) *is minimized by a possibly complex matrix $X \in C^{n \times n}$, then it is minimized by $\bar{X}$, $X + \bar{X}$, $X - \bar{X}$, $X + X^H$ and $X - X^H$ (whenever these are nonzero). In particular, if* (1) *is minimized by $X \in C^{n \times n}$ and the real part of X is not skew-symmetric, then* (1) *has a real symmetric minimizer.*

Proof. Similar to Lemma 1. □

An immediate consequence of the corollary is

THEOREM 3. *If the minimum of* (1) *is zero, then it is achieved by a symmetric matrix. Furthermore, if A is nonsingular, then the minimum is also achieved by a skew-symmetric matrix.*

Proof. If A is singular, then there is a vector $v \in \mathbf{R}^n$ such that $A^T v = 0$. The symmetric matrix $X = vv^T$ minimizes (1). If A is nonsingular and $K \neq 0$ is a skew-symmetric matrix such that $A^T K + KA = 0$, then KA is a nonzero symmetric matrix and

$$A^T(KA) + (KA)A = A^T(KA) + (-A^T K)A = 0. \tag{7}$$

Similarly, if A is nonsingular and $X \neq 0$ is a symmetric matrix such that $A^T X + XA = 0$, then XA is a nonzero skew-symmetric matrix that minimizes (1). □

The conjecture can also be proved for the class of normal matrices.

THEOREM 4. *If $A \in \mathbf{R}^{n \times n}$ is normal, then* (1) *is minimized by a symmetric matrix X.*

Proof. Since A is normal, there is a unitary matrix $U \in C^{n \times n}$ and a diagonal matrix $D \in C^{n \times n}$ such that $A = UDU^H$. The Frobenius norm is invariant under unitary transformations, so if $X \in \mathbf{R}^{n \times n}$, then

$$\frac{\|A^T X + XA\|}{\|X\|} = \frac{\|D^H(UXU^H) + (UXU^H)D\|}{\|UXU^H\|}. \tag{8}$$

Let $W = UXU^H$ and denote the diagonal entries of D by d_i. Define integers k and l by $|\bar{d}_k + d_l| = \min |\bar{d}_i + d_j|$. In terms of W and D, the right-hand side of (8) becomes

$$\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |w_{ij}(\bar{d}_i + d_j)|^2} \Big/ \|W\|$$

which is minimized by setting $W = [w_{ij}]$ where

$$w_{ij} = \begin{cases} \omega & \text{if } i = k \text{ and } j = l, \\ 0 & \text{if } i \neq k \text{ or } j \neq l, \end{cases}$$

and ω is a number of unit modulus chosen so that the real part of $U^H(W + W^H)U$ is nonzero. By Corollary 2, $(W + W^H)$ also minimizes the right-hand side of (8). So $X = U^H(W + W^H)U$ minimizes the left-hand side. Again applying Corollary 2, the real part of $U^H(W + W^H)U$ also minimizes (8). □

The conjecture holds without any special assumptions for 2-by-2 matrices.

THEOREM 5. *If A is 2-by-2, then (1) is minimized by a symmetric matrix.*

Proof. By Lemma 1, either a symmetric or a skew-symmetric matrix minimizes (1). We will show that for every 2-by-2, skew-symmetric matrix, there is a symmetric matrix that makes the quotient in (1) at least as small.

Suppose

$$A = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$$

is a real, 2-by-2 matrix. Without loss of generality, we may assume $\alpha = \delta$. To see this, let t be a root of the quadratic equation

$$(\alpha - \delta) + 2t(\gamma + \beta) - t^2(\alpha - \delta) = 0.$$

The discriminant is $4(\gamma + \beta)^2 + 4(\alpha - \delta)^2 \geqq 0$, so t is real. Set $c = (1 + t^2)^{-1/2}$ and $s = ct$. So, $c^2 + s^2 = 1$, and

$$R = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

is orthogonal. These choices of c and s make the diagonal entries of $\hat{A} = RAR^T$ equal. The Frobenius norm is unitarily invariant, so

$$\min_{X \neq 0} \frac{\|A^T X + XA\|}{\|X\|} = \min_{\hat{X} \neq 0} \frac{\|\hat{A}^T \hat{X} + \hat{X}\hat{A}\|}{\|\hat{X}\|}.$$

If the right-hand side is minimized by a symmetric matrix $\hat{X}$, then the left-hand side is minimized by the symmetric matrix $X = R\hat{X}R^T$.

All 2-by-2, skew-symmetric matrices are scalar multiples of

$$K = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

The only value of the quotient in (1) for a skew-symmetric matrix is $\|A^T K + KA\| / \|K\| = |\alpha + \delta|$. If $\alpha = \delta$, and c and s are real numbers such that $c^2 + s^2 = 1$, then for the symmetric matrix

$$X = \begin{bmatrix} c & 0 \\ 0 & s \end{bmatrix}$$

the quotient in (1) is also $\|A^T X + XA\| / \|X\| = 2|\alpha| = |\alpha + \delta|$. □

An important special case of the Lyapunov equation (2) is the case that A is stable, i.e., all eigenvalues of A have negative real part. It is well known that the Lyapunov operator is order preserving for stable matrices. A consequence is that the conjecture is true for stable matrices A. To establish this, we need the following two lemmas.

LEMMA 6. *Suppose all eigenvalues of $A \in \mathbf{R}^{n\times n}$ have negative real part and suppose B and C are (possibly complex) Hermitian matrices such that $B \geqq C$. If X and Y satisfy the Lyapunov equations $A^TX + XA = B$ and $A^TY + YA = C$ then $Y \geqq X$.*

Proof. The difference $X - Y$ satisfies the Lyapunov equation

$$A^T(X-Y)+(X-Y)A = B-C.$$

Since A is stable and $B - C$ is positive semi-definite, $X - Y$ is negative semi-definite [12, p. 277]. □

We also need

LEMMA 7. *If X and Y are (possibly complex) n-by-n Hermitian matrices such that $X \geqq Y \geqq -X$, then $\|X\| \geqq \|Y\|$.*

Proof. Let $Y = UDU^H$ be a unitary spectral-decomposition of Y (i.e. U is unitary and D is diagonal). Set $Z = UXU^H$. Note that $Z \geqq D \geqq -Z$. In particular, for $i = 1, 2, 3, \cdots n$, $z_{ii} - d_{ii} \geqq 0$ and $z_{ii} + d_{ii} \geqq 0$. So $z_{ii} \geqq |d_{ii}|$ and

$$\|Y\|^2 = \|D\|^2 = \sum_{i=1}^{n} d_{ii}^2 \leqq \sum_{i=1}^{n} z_{ii}^2 \leqq \|Z\|^2 = \|X\|^2.$$

The first and last equalities follow from the unitary invariance of the Frobenius norm.

We can now prove

THEOREM 8. *If the eigenvalues of $A \in \mathbf{R}^{n\times n}$ have negative real part, then* (1) *is minimized by a symmetric matrix.*

Proof. For a real skew-symmetric matrix $K \neq 0$, we will exhibit a symmetric matrix S (depending on K) such that

$$\frac{\|A^TK+KA\|}{\|K\|} \geqq \frac{\|A^TS+SA\|}{\|S\|}.$$

The theorem then follows from Lemma 1.

Let $K \in \mathbf{R}^{n\times n}$ be skew-symmetric and set $M = A^TK + KA$. M is skew-symmetric, so it has an orthogonal spectral-decomposition of the form $M = UDU^T$ where $U \in \mathbf{R}^{n\times n}$, $U^TU = I$, $D \in \mathbf{R}^{n\times n}$, $D = -D^T$, and $D = \operatorname{diag}(D_{jj})$ is block diagonal with 1-by-1 and 2-by-2 blocks. The skew-symmetry of D forces the 1-by-1 blocks to be zero and the 2-by-2 blocks to take the form

$$D_{jj} = \begin{bmatrix} 0 & \lambda_j \\ -\lambda_j & 0 \end{bmatrix}. \tag{9}$$

The λ's are real and (without loss of generality) positive. Define 1-by-1 and 2-by-2 matrices E_{jj} as follows. If D_{jj} is a 1-by-1 zero block, then let $E_{jj} = D_{jj}$. If D_{jj} is of the form (9), then

$$E_{jj} = \begin{bmatrix} \lambda_j & 0 \\ 0 & \lambda_j \end{bmatrix}.$$

For each j,

$$E_{jj} \geqq iD_{jj} \geqq -E_{jj}$$

where $i^2=-1$. Define $E\in \mathbf{R}^{n\times n}$ as the block diagonal matrix $E=\operatorname{diag}(E_{jj})$, and set $R=UEU^T$. The unitary invariance of the Frobenius norm shows that

$$\|R\|^2=\|E\|^2=2\sum \lambda_j^2=\|D\|^2=\|M\|^2.$$

Also $R\geqq iM\geqq -R$. If S solves

$$A^TS+SA=R,$$

Lemma 6 shows that $S\geqq iK\geqq -S$, and Lemma 7 shows $\|S\|\geqq \|iK\|=\|K\|$. Therefore

$$\frac{\|A^TS+SA\|}{\|S\|}=\frac{\|R\|}{\|S\|}=\frac{\|M\|}{\|S\|}\leqq\frac{\|M\|}{\|K\|}=\frac{\|A^TK+KA\|}{\|K\|}. \qquad \square$$

The difficulty of establishing the symmetry of the minimizer leads to the question of the symmetry of the maximizer. We close this section by showing that the maximizer of (1), the "vector" corresponding to the largest singular value of the Lyapunov operator, is symmetric.

THEOREM 9. *For all* $A\in \mathbf{R}^{n\times n}$,

$$\max_{X\neq 0}\frac{\|A^TX+XA\|}{\|X\|} \tag{10}$$

is achieved by a symmetric matrix X.

Proof. An analogue of Lemma 1 shows that the maximizer is either symmetric or skew-symmetric. So, for any skew-symmetric matrix K, it suffices to construct a nonskew-symmetric matrix S such that

$$\frac{\|A^TK+KA\|}{\|K\|}\leqq\frac{\|A^TS+SA\|}{\|S\|}. \tag{11}$$

Suppose that $n=2m$ for some integer m. (The case of n odd is similar.) There is an orthogonal matrix $U\in\mathbf{R}^{n\times n}$ and a matrix $D\in \mathbf{R}^{m\times m}$ such that

$$U^TKU=\begin{bmatrix}0 & D\\ -D^T & 0\end{bmatrix}. \tag{12}$$

This decomposition can be obtained from the Schur decomposition $D=QJQ^T$, where $Q\in\mathbf{R}^{n\times n}$ is orthogonal and $J\in\mathbf{R}^{n\times n}$ is block diagonal with m 2-by-2 blocks. If P is the permutation matrix obtained from the n-by-n identity by interchanging rows j and $m+j-1$ for $j=2,3,\cdots,m$, then $K=(QP)(P^TJP)(QP)^T$ is of the form of (12).

Now partition U^TAU into m-by-m blocks as

$$U^TAU=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}.$$

The Frobenius norm is invariant under unitary transformations, so

$$\begin{aligned}\frac{\|A^TK+KA\|^2}{\|K\|^2}&=\frac{\|U^TAUU^TKU+U^TKUU^TAU\|^2}{\|U^TKU\|^2}\\ &=\frac{\|DA_{21}-A_{21}^TD^T\|^2+\|A_{12}^TD-D^TA_{12}\|^2+2\|A_{11}^TD+DA_{22}\|^2}{2\|D\|^2}\\ &\leqq\frac{\|DA_{21}\|^2+\|A_{12}^TD\|^2+\|A_{11}^TD+DA_{22}\|^2}{\|D\|^2}\\ &=\left\|\begin{bmatrix}A_{11}^T & A_{21}^T\\ A_{12}^T & A_{22}^T\end{bmatrix}\begin{bmatrix}0 & D\\ 0 & 0\end{bmatrix}+\begin{bmatrix}0 & D\\ 0 & 0\end{bmatrix}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\right\|^2\Big/\|D\|^2.\end{aligned}$$

Inequality (11) is satisfied by $S = U^T E U$ where

$$E = \begin{bmatrix} 0 & D \\ 0 & 0 \end{bmatrix}. \qquad \square$$

It is remarkable how specific are the proofs of the above theorems. The conjecture holds for normal matrices because they are diagonalized by unitary similarity transformations. It holds when the minimum in (1) is zero because both a symmetric and a skew-symmetric minimize. It holds for 2-by-2 matrices because there is essentially only one skew-symmetric, 2-by-2 matrix. It holds for stable matrices because in the stable case the Lyapunov operator respects order.

4. Conclusion. The following counterexample shows that the conjecture is not true in general.

$$A = \begin{bmatrix} -2 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

With this choice of A, the minimum in (1) is .5034 (to four significant digits). It is achieved by the real skew-symmetric matrix

$$X \approx \begin{bmatrix} 0 & .305457 & .480347 \\ -.305457 & 0 & .491479 \\ -.480347 & -.491479 & 0 \end{bmatrix}.$$

The smallest quotient in (1) that can be obtained from a symmetric matrix X is approximately .5079.

The counterexample is not compelling: there is a symmetric matrix that makes the quotient (1) almost as small as the minimizing skew-symmetric matrix. In all counterexamples we have been able to find, there has been a symmetric matrix that produces a quotient that is no more than three times the quotient for the best skew minimizer.

To obtain a heuristic estimate of how frequently the conjecture fails, we used LINPACK [6] to run four sets of Monte Carlo studies. Each set generated upper-triangular and full matrices A of sizes 3-by-3 to 8-by-8 with nonzero entries chosen from the normal $(0, 1)$ distribution or the uniform $(-1, 1)$ distribution. The singular values and "vectors" were computed using the Kronecker product matrix for the operator (2). The results are summarized in Table 1. We know of no reason why the triangular samples produced fewer skew-symmetric minimizers than the full samples.

TABLE 1
Monte Carlo: number of skew-symmetric minimizers out of 1000 *trials.*

	Uniform $(-1, 1)$		Normal $(0, 1)$	
Order	Full	Triangular	Full	Triangular
3	14	4	3	3
4	19	8	14	5
5	20	8	13	5
6	24	8	18	6
7	22	10	22	5
8	24	5	23	5

Counterexamples are rare. The conjecture is true in the common case that the eigenvalues of A have negative real parts and it is true whenever $\operatorname{sep}(A^T, -A) = 0$. Furthermore, it appears that there is always a nearly optimal symmetric minimizer, so the conjecture is true in spirit.

The case of Theorem 8 in which the eigenvalues of A have negative real part is particularly important because it is in this form that the Lyapunov equation arises in the theory of stochastic and optimal control. The cost of the quadratic regulator problem is essentially the solution of such a Lyapunov equation [13, p. 284]. Solving the algebraic Riccati equation by the Kleinman-Newton method requires the solution of a sequence of such Lyapunov equations [13, p. 285]. The covariance matrix of a linear stochastic differential equation driven by white noise is the solution of such a Lyapunov equation [13, p. 252].

REFERENCES

[1] W. F. ARNOLD III, *Numerical solution of algebraic matrix Riccati equations*, Report NWC TP 6521, Naval Weapons Center, China Lake, CA, 1984.

[2] R. H. BARTELS AND G. W. STEWART, *A solution of the equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820-826.

[3] R. BYERS, *A* LINPACK *style condition estimator for the equation* $AX - XB^T = C$, IEEE Trans. Automat. Control, 29 (1984), pp. 926-928.

[4] ———, *Numerical condition of the algebraic Riccati equation*, Proc. 1984 Joint Summer Research Conference: Linear Algebra and Its Role in Systems Theory, 1984.

[5] A. CLINE, C. MOLER, G. W. STEWART AND J. H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368-375.

[6] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER AND G. W. STEWART, LINPACK *Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

[7] G. GOLUB, S. NASH AND C. VAN LOAN, *A Hessenberg-Schur Method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909-913.

[8] P. LANCASTER, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 544-566.

[9] G. W. STEWART, *Error and perturbation bounds associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727-764.

[10] J. M. VARAH, *On the separation of two matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 216-222.

[11] H. WIMMER AND A. ZIEBUR, *Solving the matrix equation* $\sum_{\rho=1}^{r} f_\rho(A)Xg_\rho(B) = C$, SIAM Rev., 14 (1972), pp. 318-323.

[12] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

[13] D. L. RUSSELL, *Mathematics of Finite-Dimensional Control Systems*, Marcel Dekker, New York, 1979.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

004

L-FUNCTIONS AND THEIR INVERSES*

JOHN S. MAYBEE† AND GERRY M. WIENER‡

Abstract. Using concepts from qualitative matrix theory, we introduce a class of nonlinear mappings from $\mathbb{R}^n \to \mathbb{R}^n$ called L-functions. These generalize the L-matrices in much the same way that M-functions generalize M-matrices. We prove some global inverse function theorems for L-functions on several different types of domains without assuming that such functions are differentiable. Thus we do not make use of the Jacobian matrix. We also obtain interesting qualitative relations which must hold between an L-function and its inverse. Finally we prove a global implicit function theorem for L-functions, again without assuming differentiability.

Key words. Jacobians, inverse function theorems, implicit function theorems, global univalence

AMS(MOS) subject classifications. 26B10, 15A99

1. Introduction. The classical inverse function theorem and implicit function theorem are strictly local results. Therefore they are of limited utility to research in various applied fields such as economics, engineering, or regional and urban planning where nonlinear models are often formulated. In these fields scientists would like to know when the models they have built have solutions wherever data is given in the range of the functions defining the model. Thus global inverse function theorems are often required.

In recent years several global theorems have been proved. The main result seems to be the work of Gale and Nikaido [2] who deal with the so-called P-functions, i.e., functions whose Jacobian matrix is a P-matrix throughout a suitable domain. (The matrix A is a P-matrix if every principal minor of A is positive.) The monograph of Parthasarathy [11] summarizes this and other known results on global univalence although it says very little about implicit functions.

From Jacobi's time down to the present, researchers have concentrated upon the use of properties of the Jacobian matrix in order to derive inverse function theorems. We shall show that differentiability throughout a domain is not essential in order to prove either a global inverse function theorem or implicit function theorem. Thus it is not always necessary to formulate such results in terms of the Jacobian matrix.

During the past 25 years qualitative matrix theory has been developed and applied to a variety of problems such as stability (see Jeffries, Klee and van den Driessche [4]), the solution of linear systems (see [9], [8], [6], among others) controllability (see Jeffries [3]), etc. We borrow the basic ideas from this field and a fundamental concept due to Ortega and Rheinboldt [10] in order to define the concept of a qualitative mapping. From among the qualitative mappings we identify a subclass which we call L-functions and which generalize to the nonlinear case the notion of an L-matrix introduced by Klee, Ladner and Manber [6]. Our definition generalizes L-matrices in much the same way as the definition of an M-function (see, for example, Rheinboldt [12]) generalizes the notion of an M-matrix. For the L-functions we prove both global inverse function theorems and a global implicit function theorem.

2. Fundamental concepts. Given a real matrix A, we can associate with it a new matrix $Q(A)$ such that $Q(A)_{ij} = \text{sgn } a_{ij}$ for all i and j (see Maybee and Quirk [9], where this concept was first introduced and elaborated upon). For example, if

* Received by the editors August 19, 1985; accepted for publication (in revised form) April 21, 1986.

† University of Colorado, Boulder, Colorado 80309. The work of this author was supported in part by grant no. 101/84 from the Division of Scientific Affairs of the North Atlantic Treaty Organization.

‡ University of Colorado, Denver, Colorado 80202.

$$A_0 = \begin{bmatrix} -3 & 4 & 2 \\ 1 & 1 & -8 \\ 6 & -4 & 0 \end{bmatrix},$$

then

$$\operatorname{sgn} A_0 = \begin{bmatrix} - & + & + \\ + & + & - \\ + & - & 0 \end{bmatrix}.$$

Note the use of + in place of 1 and − in place of −1 here. This usage has become conventional (see Johnson [5]).

Equivalently we can associate with A a signed digraph $S(A) = (V, A, \sigma)$ where V consists of n points (vertices) labeled $1, 2, \cdots, n$, $(i, j) \in A$ is an arc of $S(A)$ if and only if $a_{ij} \neq 0$ and $\sigma\colon A \to (+, -)$ following the rule that $\sigma(i, j) = +$ if $a_{ij} > 0$ and $\sigma(i, j) = -$ if $a_{ij} < 0$. For the above matrix A_0, $S(A_0)$ is shown in Fig. 1. Note the use of a dashed line for a negative arc.

DEFINITION 1. A real $n \times n$ matrix A is an L-matrix (or sign nonsingular matrix) if and only if A is nonsingular and for all $n \times n$ matrices B satisfying $Q(B) = Q(A)$, B is also nonsingular.

In other words, A is an L-matrix if it is nonsingular by virtue of its sign pattern alone.

We will denote by $\tilde{S}(A)$ the signed digraph obtained from $S(A)$ by deleting all loops of $S(A)$. Also, as is conventional, the sign of any subset A_0 of A is simply the product of the sign of the arcs in A_0 (sign $A_0 = +$ if $A_0 = \emptyset$). The following result is the fundamental theorem on square L-matrices.

THEOREM A. (Bassett, Maybee and Quirk [1]). *A real $n \times n$ matrix A is an L-matrix if and only if by column permutations and/or multiplication of columns by* -1, *it can be transformed into a matrix B satisfying*

(i) $b_{ii} < 0$, $1 \leqq i \leqq n$,

(ii) $\tilde{S}(B)$ *has only negative cycles.*

An L-matrix B is said to be in *normal form* when it satisfies (i) and (ii) of Theorem A.

We now have the tools to introduce the concept of a qualitative function. To this end assume $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ and that $f = (f_1, \cdots, f_n)$. Let e^j be the jth standard basis vector for $\mathbb{R}^n$. Following Ortega and Rheinboldt suppose $x \in D$ and define

$$\phi_{ij}(t) = f_i(x + te^j)$$

for all t such that $x + te^j \in D$. Suppose that for all i, j, $1 \leqq i, j \leqq n$, ϕ_{ij} is either a strictly increasing, strictly decreasing, or constant function of t independent of the base point x.

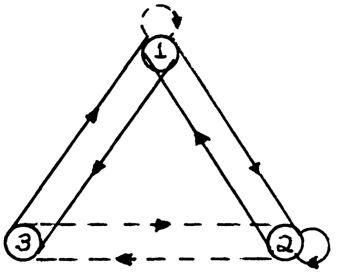


FIG. 1. *The signed digraph of* A_0.

In this case we say that f is *qualitatively defined on D* and we associate with f a matrix $Q(f)$ such that

$$Q(f)_{ij} = + \quad \text{if } \phi_{ij}(t) \text{ is a strictly increasing function,}$$

$$Q(f)_{ij} = - \quad \text{if } \phi_{ij}(t) \text{ is a strictly decreasing function, and}$$

$$Q(f)_{ij} = 0 \quad \text{if } \phi_{ij}(t) \text{ is a constant function.}$$

Thus we see that $Q(f)_{ij} = 0$ if the variable x_j "does not appear in f_i", $Q(f)_{ij} = +$ if f_i is an increasing function of x_j in D when all other variables are held constant, and $Q(f)_{ij} = -$ in a decreasing function of x_j in D when all other variables are held constant.

DEFINITION 2. The function f: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is called an *L-function on D* if f is qualitatively defined on D and $Q(f)$ is an L-matrix.

Now Theorem A can be extended to the nonlinear case because column interchanges in the matrix $Q(f)$ correspond to relabeling pairs of variables in the function f and multiplication of a column of $Q(f)$ by -1 corresponds to replacing a variable in f by its negative. Thus we can use the test given in Theorem A to determine when f is an L-function by applying the theorem to $Q(f)$. We wish to point out two facts of importance, however. The first is that at the present time it is unknown whether or not the problem of testing a signed digraph to determine that it has only negative cycles is NP-complete. Thus the problem of testing a given qualitative function to determine if it is or is not an L-function is of unknown difficulty. The second fact of importance is that it is not always desirable to put an L-matrix or an L-function into the same normal form.

Here are some examples of L-functions.

Example 1. Let $f(x_1, x_2) = (f_1(x_1, x_2), f_2(x_1, x_2))$ where

$$f_1(x_1, x_2) = \frac{a_{11}x_1 + a_{12}x_2}{a_{21}x_1 + a_{22}x_2}, \qquad a_{11}a_{22} - a_{12}a_{21} < 0,$$

$$f_2(x_1, x_2) = \frac{1}{a_{31}x_1 + a_{32}x_2} \quad \text{where } a_{ij} > 0, \quad i = 1, 2, 3, \quad j = 1, 2.$$

This is an L-function for all $x_1 > 0$ and $x_2 > 0$.

Example 2. Let $f(x_1, \cdots, x_n) = (f_1(x_1, \cdots, x_n), \cdots, f_n(x_1, \cdots, x_n))$ where

$$f_1(x_1, \cdots, x_n) = \frac{x_2 + c_1}{b_1 + a_{11}x_1},$$

$$f_i(x_1, \cdots, x_n) = \frac{x_{i+1} + c_i}{a_{i,i-1}x_{i-1} + a_{ii}x_i + b_i}, \qquad 2 \leqq i \leqq n-1,$$

$$f_n(x_1, \cdots, x_n) = \frac{1}{a_{n,n-1}x_{n-1} + a_{nn}x_n + b_n},$$

with all constants positive. This is an L-function for all $x_i \geqq 0$, $1 \leqq i \leqq n$. This example is of particular interest because the closely related example defined by

$$f_1(x_1, \cdots, x_n) = a_{11}x_1 + a_{12}x_2 + c,$$

$$f_i(x_1, \cdots, x_n) = \frac{a_{ii}x_i + a_{i,i+1}x_{i+1} + c_i}{a_{i,i-1}x_{i-1} + b_i}, \qquad 2 \leqq i \leqq n-1,$$

$$f_n(x_1, \cdots, x_n) = \frac{x_n + c_n}{a_{n,n-1}x_{n-1} + b_n},$$

with all constants positive is also an L-function for all $x_i \geqq 0$. This second case is not in the normal form required in Theorem A.

We conclude this section with two technical ideas required in the following theorems. First we say that two vectors $x, y \in \mathbb{R}^n$ *conform* in sign if for $1 \leqq i \leqq n$, $\operatorname{sgn} x_i \neq 0$ and $\operatorname{sgn} y_i \neq 0$ imply $\operatorname{sgn} x_i = \operatorname{sgn} y_i$, and for at least one i, $1 \leqq i \leqq n$, $\operatorname{sgn} x_i \neq 0$ and $\operatorname{sgn} y_i \neq 0$. The vector $x, y \in \mathbb{R}^n$ *anticonform* in sign if x and $-y$ conform in sign. Two vectors $x, y \in \mathbb{R}^n$ *nonconform* in sign if x and y neither conform nor anticonform in sign. Two vectors $x, y \in \mathbb{R}^n$ *strictly nonconform* in sign if x and y nonconform in sign and for at least one value of i $\operatorname{sgn} x_i \neq 0$ and $\operatorname{sgn} y_i \neq 0$. To illustrate observe that:

$(1, 0, 2, 0)$ and $(2, 1, 0, 1)$ conform in sign,
$(1, 2, 3, -1)$ and $(-1, -2, -3, 0)$ anticonform in sign,
$(1, 0, 5, 0)$ and $(0, 1, 0, -3)$ nonconform but do not strictly nonconform in sign,
$(1, 3, 5)$ and $(2, -1, 2)$ strictly nonconform in sign.

DEFINITION 3. A domain $D \subseteq \mathbb{R}^n$ will be called *coordinately connected* if, given any two points $a, b \in D$, there exists a finite sequence of distinct points $p_1 = a, p_2, \cdots, p_k = b$ in D such that

(A) each vector $p_{i+1} - p_i$, $i = 1, \cdots, k-1$, has exactly the same sign pattern, i.e., $\operatorname{sgn}(p_{i+1} - p_i) = \operatorname{sgn}(b - a)$, and

(B) there is a path from p_i to p_{i+1}, along the edges of the parallelepiped enclosed by p_i, p_{i+1} for $i = 1, \cdots, k-1$ which lies entirely within D.

Note, for example, that D_1 in Fig. 2 is coordinately connected but D_2 is not because each path from a to b violates condition (A).

3. Univalence. For reference we state the classical inverse function theorem. In this connection we shall use $J(f)$ to denote the Jacobian matrix of the function $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^n$. The value of this matrix valued function at the point x is $J(f)(x)$.

THEOREM B. *Suppose $f \in C^1(D)$ on the open set $D \subseteq \mathbb{R}^n$ into $\mathbb{R}^n$ and $[J(f)(a)]^{-1}$ exists for some $a \in D$. Then*

(a) *there exist open sets U, V in $\mathbb{R}^n$ with $a \in U$, $f(a) \in V$ and f is a univalent mapping of U onto V, and*

(b) *$f^{-1}\colon V \to U$ is $C^1(V)$ and for all $y \in V$*

$$J(f^{-1})(y) = [J(f)(y)]^{-1}.$$

If $J(f)(x)$ exists for all $x \in D$, it is not necessarily true that f is globally univalent as can be seen by the example $f(x, y) = (e^x \cos y, e^x \sin y)$. We have

$$|J(f)(x,y)| = \begin{vmatrix} e^x \cos y & -e^x \sin y \\ e^x \sin y & e^x \cos y \end{vmatrix} = e^{2x} \neq 0$$

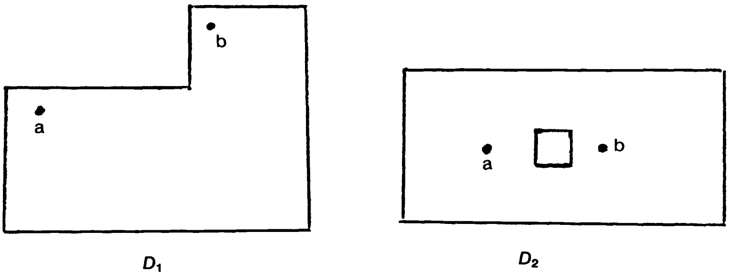


FIG. 2. *D_1 is coordinately connected, D_2 is not.*

for all $(x, y) \in \mathbb{R}^2$, but, since $f(x, y + 2\pi) = f(x, y)$, f is not globally univalent on all of $\mathbb{R}^2$. Note, on the other hand, that $Q(f)$ is an L-matrix on $\mathbb{R} \times (0, \pi/2)$, $\mathbb{R} \times (\pi/2, \pi)$, etc. Our first theorem shows that f is, in fact, globally univalent on each of these semi-infinite rectangles.

THEOREM 1. *Let f: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be an L-function on an open or closed rectangle D. Then f is globally univalent on D.*

Proof. Suppose $a, b \in D$, $a \neq b$ with $a = (a_1, a_2, \cdots, a_n)$, $b = (b_1, b_2, \cdots, b_n)$. Then

$$\begin{aligned} f_i(b)-f_i(a) = {} & f_i(b_1, a_2, \cdots, a_n) - f_i(a_1, a_2, \cdots, a_n) + f_i(b_1, b_2, a_3, \cdots, a_n) \\ & - f_i(b_1, a_2, \cdots, a_n) + \cdots + f_i(b_1, b_2, \cdots, b_n) \\ & - f_i(b_1, b_2, \cdots, b_{n-1}, a_n) \end{aligned}$$

for $i = 1, 2, \cdots, n$. Observe that $(a_1, a_2, \cdots, a_n)$, $(b_1, a_2, \cdots, a_n)$, $\cdots$, $(b_1, b_2, \cdots, b_n)$ are all points in D since D is rectangular. If $(b - a)$ and row$_i$ $Q(f)$ conform or anticonform in sign, then

$$\operatorname{sgn}(f_i(b) - f_i(a)) = \pm \operatorname{sgn} \text{row}_i\, Q(f) \cdot (b - a) \neq 0.$$

Here $\cdot$ represents the standard scalar product. Since $Q(f)$ is an L-matrix and $(b - a)$ is not the zero vector, there exists a row of $Q(f)$, say row$_c$ $Q(f)$, such that row$_c$ $Q(f)$ and $b - a$ either conform or anticonform in sign. Thus $f_c(b) - f_c(a) \neq 0$ implying $f(b) - f(a) \neq 0$. Since a and b were arbitrary, distinct points in D, f must be globally univalent on D. □

Note that D is not required to be a finite rectangle in the proof so the assertion made above regarding the function $f(x, y) = (e^x \cos y, e^x \sin y)$ on domains such as $\mathbb{R} \times (0, \pi/2)$, etc., are correct.

Our next result shows how the domain, D, can be enlarged.

THEOREM 2. *Let f: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be an L-function on a coordinately connected domain D. Then f is globally univalent on D.*

Proof. Suppose for contradiction that f is not globally invertible, i.e., there exist points $a, b \in D$, $a \neq b$, such that $f(a) = f(b)$. Since D is coordinately connected, there exists a sequence of points $p_1, p_2, \cdots, p_k$ in D with $p_1 = a$, $p_k = b$ such that (A) and (B) are satisfied. Now (B) implies that for each pair p_j, p_{j+1} there is a sequence of n points $q_{1j}, q_{2j}, \cdots, q_{nj}$ such that $q_{ij} \in D$, $1 \leqq i \leqq n$, $q_{1j} = p_j$, $q_{nj} = p_{j+1}$ and q_{ij}, $q_{i+1,j}$ differ in at most one coordinate. ($q_{1j}, \cdots, q_{nj}$ will be equal to the vertices of the parallelepiped that lie on the path from p_j to p_{j+1}. If the parallelepiped induced by p_j and p_{j+1} has dimension $m \leqq n$, then exactly m of $q_{1j}, q_{2j}, \cdots, q_{nj}$ will be distinct.) Thus for all i, $1 \leqq i \leqq n$,

$$f_i(p_{j+1}) - f_i(p_j) = f_i(q_{2j}) - f_i(q_{1j}) + f_i(q_{3j}) - f_i(q_{2j}) + \cdots + f_i(q_{nj}) - f_i(q_{n-1,j}).$$

As in the proof of Theorem 1, if $p_{j+1} - p_j$ and row$_i$ $Q(f)$ conform or anticonform in sign, then

$$\operatorname{sgn}(f_i(p_{j+1}) - f_i(p_j)) = \operatorname{sgn} \text{row}_i\, Q(f) \cdot (p_{j+1} - p_j) \neq 0.$$

Since $Q(f)$ is an L-matrix and $p_{j+1} - p_j$ is never the zero vector, there exists a row of $Q(f)$, say row$_c$ $Q(f)$, such that row$_c$ $Q(f)$ and $p_{j+1} - p_j$ either conform or anticonform in sign. Moreover, by condition (A) in the definition of coordinate connectedness, each $p_{j+1} - p_j$, $j = 1, \cdots, k - 1$, has the same sign pattern so row$_c$ $Q(f)$ and $p_{j+1} - p_j$ will either conform or anticonform in sign for each j. Consequently either $f_c(p_{j+1}) - f_c(p_j) > 0$ for $1 \leqq j \leqq k - 1$ or $f_c(p_{j+1}) - f_c(p_j) < 0$ for $1 \leqq j \leqq k - 1$ (the difference being positive when $\text{row}_c\, Q(f)$ and $p_{j+1} - p_j$ conform in sign and negative when $\text{row}_c\, Q(f)$ and $p_{j+1} - p_j$ anticonform in sign). Since

$$f_c(b) - f_c(a) = f_c(p_k) - f_c(p_{k-1}) + f_c(p_{k-1}) - f_c(p_{k-2}) + \cdots + f_c(p_2) - f_c(p_1),$$

we have $f_c(b) - f_c(a) \neq 0$, when $p_{j+1} - p_j$ and $\text{row}_c\, Q(f)$ conform or anticonform in sign. But this implies $f_c(b) - f_c(a) \neq 0$ when $b - a$ and $\text{row}_c\, Q(f)$ conform or anticonform in sign as $b - a$ and $p_{j+1} - p_j$ have the same sign pattern. The remainder of the proof now follows the proof of Theorem 1. □

The following result is now of some interest.

COROLLARY 3. *If f is an L-function on an open convex region D, then f is globally univalent on D.*

Proof. Suppose a, b are distinct points of D. Since the line segment $[a, b]$ is compact, there exists a radius $2r$ such that for each $p \in [a, b]$, the open ball $B(p, 2r)$ with center p and radius $2r$ lies entirely within D. If p, q are on the line ab and the distance from p to q is less than or equal to r, then it follows that the closed cube with diagonal pq must lie entirely within D. It is thus clear that there exists $p_1, p_2, \cdots, p_k$ on $[a, b]$ such that properties (A) and (B) are satisfied. Thus Theorem 2 applies to D. □

As an example illustrating Theorem 2 consider the following function discussed by Gale and Nikaido, namely

$$f(x, y) = e^{2x} - y^2 + 3, \qquad g(x, y) = 4e^{2x}y - y^3.$$

Observe that when $x = 0$, $y = \pm 2$, we have $f = g = 0$. Now (f, g) is an L-function for all x and for $0 < y < 2e^x/\sqrt{3}$. Also this region is coordinately connected although not convex, hence the mapping defined by f and g is globally univalent there.

We can extend Theorem 2 and Corollary 3 by using the invariance of the domain theorem which states that if D is open in $\mathbb{R}^n$ and f: $D \to \mathbb{R}^n$ is globally univalent and continuous, then $f(D)$ is open in $\mathbb{R}^n$ and f is a homeomorphism [11]. We therefore have the following result.

THEOREM 4. *If f: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a continuous L-function and D is an open coordinately connected set or if D is a compact coordinately connected set, then f is a homeomorphism.*

It should be noted that if $f \in C^1(D)$, then the usual inverse function theorem and its related results apply so we obtain also the following results.

THEOREM 5. *If f: $D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a continuously differentiable L-function and D is an open coordinately connected set, then f^{-1} is continuously differentiable on $f(D)$.*

We will present here two applications of our results thus far. In this connection we point out that our definitions imply that L-functions are normalized so that each $\phi_{ii}(t)$, $i = 1, \cdots, n$ is different from zero. The usual normalization is that given by Bassett, Maybee and Quirk [1] where $\phi_{ii}(t)$ is strictly decreasing, i.e., $Q(f)_{ii} = -$, $i = 1, \cdots, n$.

As our first application suppose $f(z)$ in an analytic function of the complex variable z on the convex or coordinately connected domain $D \subseteq \mathbb{R}^2$. Then in $D\, f(z) = u(x, y) + iv(x, y)$ satisfies the Cauchy Riemann equations $u_x = v_y$, $u_y = -v_x$. Assume u_x and u_y do not change sign in D and that at least one of u_x, u_y is nonzero in D. Then the Jacobian matrix of the mapping $f = (u, v)$ is

$$J(f)(x, y) = \begin{bmatrix} u_x & u_y \\ -u_y & u_x \end{bmatrix}.$$

If $u_x > 0$ in D, then this matrix is an L-matrix throughout D and

$$Q(f) = \begin{bmatrix} + & + \\ - & + \end{bmatrix} \text{ for } u_y > 0, \quad \begin{bmatrix} + & - \\ + & + \end{bmatrix} \text{ for } u_y < 0, \quad \text{or} \begin{bmatrix} + & 0 \\ 0 & + \end{bmatrix} \text{ for } u_y = 0$$

in D. If $u_x < 0$ in D, then the matrix is again an L-matrix throughout D with

$$Q(f)=\begin{bmatrix} - & + \\ - & - \end{bmatrix} \text{ for } u_y>0, \quad \begin{bmatrix} - & - \\ + & - \end{bmatrix} \text{ for } u_y<0, \quad \text{or} \begin{bmatrix} - & 0 \\ 0 & - \end{bmatrix} \text{ for } u_y=0$$

in D. Similar results hold for $u_x = 0$, $u_y \neq 0$ in D. Thus in all cases we can assert that if $f(x)$ is analytic in D with u_x and u_y of constant sign at least one of which is nonzero, then the mapping (u, v) defines an L-function on D. Therefore by our theorems (u, v) or, equivalently, f is globally univalent on D.

As a second application we consider what form an L-matrix A takes when the normalization is $a_{ii} > 0$, $i = 1, 2, \cdots, m$. We then have that A is an L-matrix if and only if every cycle of even length in $\tilde{S}(A)$ is negative and every cycle of odd length positive. For such a matrix every principal minor of order r is positive for all r, i.e., A is a P-matrix. Thus we can identify the class of qualitative P-matrices, a subclass of P-matrices apparently not noticed before. It follows that there is a subclass of P-functions, the qualitative P-functions, which are globally univalent as a consequence of our theorems.

In his book Parthasarathy presents an example, namely, $f(x, y, z) = (f_1, f_2, f_3)$ where $f_1 = x^2 + z^2, f_2 = x^2 + y^2, f_3 = y^2 + z^2$. The Jacobian is

$$J(f)(x,y,z)=\begin{bmatrix} 2x & 0 & 2z \\ 2x & 2y & 0 \\ 0 & 2y & 2z \end{bmatrix}.$$

It is easy to see that for all $x < 0$, $y < 0$, $z < 0$ this is an L-matrix as it is also for all $x > 0$, $y > 0$, $z > 0$. On the other hand, for all $x > 0$, $y > 0$, $z > 0$, it is also a P-matrix and

$$Q(f)=\begin{bmatrix} + & 0 & + \\ + & + & 0 \\ 0 & + & + \end{bmatrix}$$

is a qualitative P-matrix. Note that the related function $f_1 = x^2 - z^2, f_2 = y^2 - x^2, f_3 = z^2 + y^2$ has

$$Q(f)=\begin{bmatrix} + & 0 & - \\ - & + & 0 \\ 0 & + & + \end{bmatrix}$$

and $J(f)(x, y, z)$ is also a P-matrix for all $x > 0$, $y > 0$, $z > 0$.

We could, of course, generate families of additional examples.

4. The sign pattern of the inverse. The classical inverse function theorem (Theorem B) says something about the Jacobian matrix of the inverse mapping. On the other hand, the following result of Lady and Maybee [7] details the sign pattern of the inverse of an L-matrix.

THEOREM C. (Lady and Maybee). *Let A be an irreducible L-matrix with $a_{ii} \neq 0$, $1 \leqq i \leqq n$. Then, setting $A^{-1} = [\alpha_{ij}]$, we have*

(i) *if $a_{ij} \neq 0$*, sgn α_{ji} = sgn a_{ij},

(ii) *if $a_{ij} = 0$, then the sign of α_{ji} is qualitatively determined if and only if every path $p(j \twoheadrightarrow i)$ in $\tilde{S}(A)$ has the same sign. In this case,* sgn α_{ji} = $-$sgn $p(j \twoheadrightarrow i)$ *where $p(j \twoheadrightarrow i)$ is any path in $\tilde{S}(A)$ from j to i.*

For example, suppose

$$Q(A)=\begin{bmatrix} - & + & 0 & + \\ - & - & - & 0 \\ 0 & + & - & - \\ - & 0 & + & - \end{bmatrix},$$

then

$$\operatorname{sgn} A^{-1} = \begin{bmatrix} - & - & * & - \\ + & - & + & * \\ * & - & - & + \\ + & * & - & - \end{bmatrix}$$

where we use $*$ to denote an element whose sign is not determinate because of paths of different signs from j to i in $\tilde{S}(A)$.

Theorem C shows that the structure of the inverse of an irreducible L-matrix is to a large degree qualitatively determined. One would hope that a similar result holds for L-functions. In fact we can say even more.

THEOREM 6. *Suppose f: $D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an L-function and D is coordinately connected. If* sgn $(Q(f))^{-1}_{i,j}$ *is qualitatively determined, then*:

(α) sgn $(Q(f))^{-1}_{i,j} = +$ *implies $f_i^{-1}(x + te^j)$ is increasing with respect to t for any $x \in f(D)$.*

(β) sgn $(Q(f))^{-1}_{i,j} = 0$ *implies $f_i^{-1}(x + te^j)$ is constant with respect to t for any $x \in f(D)$.*

(γ) sgn $(Q(f))^{-1}_{i,j} = -1$ *implies $f_i^{-1}(x + te^j)$ is decreasing with respect to t for any $x \in f(D)$.*

Proof. Suppose $(Q(f))^{-1}_{i,j}$ is qualitatively determined. Let $f^{-1}(x) = u, f^{-1}(x + te^j) = v$ for some $t > 0$. Thus $f(u) = x, f(v) = x + te^j$ and $f(v) - f(u) = te^j$. We now claim that for the above j,

(1) row$_j$ $(Q(f))$ and $v - u$ conform in sign,

(2) if $i \neq j$, row$_i$ $(Q(f))$ and $v - u$ nonconform in sign.

Facts (1) and (2) follow from the proof of Theorem 2 in which we found out that

$$\operatorname{sgn}(f_i(v) - f_i(u)) = \pm \operatorname{sgn} \operatorname{row}_i Q(f) \cdot (v - u) \neq 0$$

if row$_i$ $Q(f)$ and $v - u$ conform or anticonform in sign. If row$_j$ $Q(f)$ and $v - u$ neither conform nor anticonform in sign, there must be another row in $Q(f)$, say row$_c$ $Q(f)$, such that row$_c$ $Q(f)$ and $v - u$ either conform or anticonform in sign. This is due to the fact that $Q(f)$ is an L-matrix. But then $f_c(v) - f_c(u) \neq 0$ for some $c \neq j$, contradicting the fact that $f(v) - f(u) = f_j(v) - f_j(u) = te^j$. Since row$_j$ $Q(f)$ and $v - u$ cannot anticonform in sign as $t > 0$, it is clear that (1) and (2) hold.

We further claim that there exists an L-matrix A having identical sign patterns of $Q(f)$ such that $A(v - u) = e^j$. To prove this, let $v - u = (\omega_1, \omega_2, \cdots, \omega_n)$. Define A as follows:

$$A_{ik} = \operatorname{sgn} Q(f)_{ik} \quad \text{if } \omega_k = 0,$$

$$A_{ik} = 0 \quad \text{if } Q(f)_{ik} = 0,$$

$$A_{ik} = \frac{1}{c_i \omega_k} \quad \text{when } Q(f)_{ik} \text{ and } \omega_k \text{ have the same nonzero sign.}$$

(Here c_i equals the number of corresponding elements in row$_i$ $Q(f)$ and ω having the same nonzero sign.)

$$A_{ik} = \frac{-1}{d_i \omega_k} \quad \text{when } Q(f)_{ik} \text{ and } \omega_k \text{ have opposite signs.}$$

(Here d_i equals the number of corresponding elements in row$_i$ $Q(f)$ and ω that have opposite signs.) From properties (1) and (2) and the definition of A it is evident that A is an L-matrix.

Since there exists an L-matrix A having the same sign pattern as $Q(f)$ such that $A\omega = e^j$, any qualitative determination found for elements in $\operatorname{col}_j (Q(f))^{-1}$ must hold for ω and vice versa. Thus, if $\operatorname{sgn} (Q(f))_{ij}^{-1}$ is qualitatively determined, $\operatorname{sgn} \omega_i = \operatorname{sgn} (Q(f))_{ij}^{-1}$. But $\omega_i = f_i^{-1}(x + te^j) - f_i^{-1}(x)$ and x was arbitrary, so that properties (α), (β), and (γ) must hold. □

From the proof of Theorem 6 we obtain the following interesting corollary.

COROLLARY 7. *If A is an L-matrix, then* $\operatorname{row}_j A$ *and* $\operatorname{col}_j A^{-1}$ *conform in sign for* $1 \leqq j \leqq n$, *and* $\operatorname{row}_j A$ *and* $\operatorname{col}_k A^{-1}$ *nonconform in sign for* $1 \leqq j, k \leqq n, j \neq k$.

5. An implicit function theorem. Before presenting an implicit function theorem for L-functions, we will present two general implicit function theorems that apply even for nondifferentiable functions.

THEOREM 8. *Suppose $f\colon D \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and that there exists a point $(a, b) \in D$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ such that $f(a, b) = 0$. Define $F\colon D \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ by $F(x, y) = (x, f(x, y))$. If F is globally univalent, then the set $S = \{(x, y) \in D, x \in \mathbb{R}^n, y \in \mathbb{R}^m\colon f(x, y) = 0\}$ defines a nontrivial implicit function $y = h(x)$, i.e., if $(x, y) \in S$ and $(x, z) \in S$, then $y = z = h(x)$.*

Proof. Suppose $f(x, y) = 0$ and $f(x, z) = 0$ where $(x, y) \in D$, $(x, z) \in D$, $x \in \mathbb{R}^n$, $y, z \in \mathbb{R}^m$. Since $(x, f(x, y)) = (x, f(x, z))$, we have $F(x, y) = F(x, z)$ and thus $y = z$ as F is one-to-one. □

THEOREM 9. *Suppose $f\colon D \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and there exists a point $(a, b) \in D$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ such that $f(a, b) = 0$. Suppose N is a neighborhood of (a, b) contained entirely within D. If F as defined in Theorem 8 is a homeomorphism, then the set $S = \{(x, y) \in D, x \in \mathbb{R}^n, y \in \mathbb{R}^m\colon f(x, y) = 0\}$ defines a continuous, nontrivial implicit function $y = h(x)$ and there exists an open set $H \subseteq \mathbb{R}^n$ such that $a \in H \subseteq$ domain h.*

Proof. Let U, V be open sets containing a, b, respectively, such that $U \times V \subseteq N$. By the invariance of domain theorem, $F(U \times V)$ is an open set containing $(a, 0)$. Thus there exist open sets H, K containing a, 0, respectively, such that $H \times \{0\} \subseteq H \times K \subseteq F(U \times V)$. Clearly, $H \subseteq U$ by the definition of F. The remainder of the theorem follows from Theorem 8. □

We will now present an implicit function theorem for L-functions.

THEOREM 10. *Suppose $f\colon D \subseteq \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ and there exists a point $(a, b) \in D$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, such that $f(a, b) = 0$. Suppose D is coordinately connected and that $f(x, y)$ is an L-function with respect to y for all x, $(x, y) \in D$. Then F as defined in Theorem 8 is an L-function on D and Theorem 8 holds.*

If in addition f is continuous, D is compact and $(a, b) \in N \subseteq D$ where N is a neighborhood of (a, b) or if f is continuous and D is open, then Theorem 9 holds.

Proof. Since f is an L-function with respect to y for all x for $(x, y) \in D$ we can introduce the matrix $Q(F)$ defined as follows:

$$Q(F) = \begin{bmatrix} q_{11} & \vdots & 0 \\ 0 & \vdots & \\ & q_{nn} & \\ q_{n+1,1} & & q_{n+1,n+m} \\ \cdots & & \\ q_{n+m,1} & & q_{n+m,n+m} \end{bmatrix}.$$

Here the elements $q_{ii} = +$, $1 \leqq i \leqq n$, and the elements q_{ij}, $i = n + 1, \cdots, n + m$, $j = n + 1, \cdots, n + m$, are determined by the fact that f is an L-function with respect to y. Thus $Q(F)$ is block lower triangular and the diagonal blocks are L-matrices. By an easy argument $Q(F)$ is then an L-matrix regardless of the numbers Q_{ij}, $i = n + 1, \cdots, n + m$, $j = 1, \cdots, n$, which we may arbitrarily assign the values $+$, for example. The theorem now follows from Theorems 2, 8 and 9. □

Theorem 6 may now be employed to provide us with information concerning the implicit function $y = h(x)$. In this regard assume that $h(x) = (h_1(x), h_2(x), \cdots, h_m(x))$ and that $F^{-1}(x, y) = (g_1(x, y), g_2(x, y), \cdots, g_{n+m}(x, y))$. Since $y = h(x)$ iff $(x, y) = F^{-1}(x, 0)$,

$$h_i(x+te^j)-h_i(x)=g_{n+i}(x+te^j,0)-g_{n+i}(x,0).$$

Thus if $\operatorname{sgn}(Q(F))^{-1}_{n+i,j}$ is qualitatively determined, we can determine if $h_i(x + te^j) - h_i(x)$ is increasing, decreasing or constant with respect to t.

REFERENCES

[1] L. Bassett, J. Maybee and J. Quirk, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 26 (1968), pp. 544–563.

[2] D. Gale and H. Nikaido, *The Jacobian matrix and global univalence of mappings*, Math. Ann., 159 (1965), pp. 81–93.

[3] J. Jeffries, *Sign stable and sign controllable systems*, submitted.

[4] C. Jeffries, V. Klee and P. van den Driessche, *When is a matrix sign stable?*, Canad. J. Math., 29 (1977), pp. 315–326.

[5] C. Johnson, *Combinatorial Matrix Theory*, The Johns Hopkins University Press, Baltimore, MD, to be published.

[6] V. Klee, R. Ladner and R. Manber, *Signsolvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.

[7] G. Lady and J. Maybee, *Qualitatively invertible matrices*, J. Math. Social Sciences, 6 (1983), pp. 397–407.

[8] J. Maybee, *Sign solvability*, in Computer Assisted Analysis and Model Simplification, H. J. Greenberg and J. S. Maybee, eds., Academic Press, New York, 1981.

[9] J. Maybee and J. Quirk, *Qualitative problems in matrix theory*, SIAM Rev., 11 (1969), pp. 30–51.

[10] J. Ortega and W. Rheinboldt, *On a class of approximate iterative processes*, Arch. Rational Mech. Anal., 23 (1967), pp. 352–365.

[11] T. Parthasarathy, *On Global Univalence Theorems*, Lecture Notes 977, Springer-Verlag, Berlin, Heidelberg, New York, 1983.

[12] W. Rheinboldt, *On M-functions and their application to nonlinear Gauss–Seidel iterations and to network flows*, J. Math. Ann. Appl., 32 (1970), pp. 274–307.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

THRESHOLD REPRESENTATIONS OF MULTIPLE SEMIORDERS*

JEAN-PAUL DOIGNON†

Abstract. Cozzens and Roberts recently proved a variant of the Scott and Suppes representation theorem for semiorders. They treated nested pairs of semiorders, but stated as open the corresponding problem with any number of relations. A solution is described which even alleviates the condition and the proof in the case of two relations. Moreover, representations simultaneously involving constant and nonconstant thresholds are considered.

Key words. semiorder, interval order, threshold, preference modelling

AMS(MOS) subject classifications. 06A10, 05C20, 92A25, 90A10

1. Introduction. In order to modelize an individual's preferences on a finite set X of objects or actions, Luce [5] associates to every object x in X a measure of *utility* $f(x)$, and proposes to explain preferences through comparisons of utility values, taking into account a discrimination *threshold* σ. Formally a *semiorder* is a binary relation P on a finite set X which satisfies the conditions in the following proposition from Scott and Suppes [14].

PROPOSITION 1. *There exist a real-valued mapping f on X, and a nonnegative real number σ such that for all $x, y \in X$,*

$$xPy \Leftrightarrow f(x) > f(y) + \sigma$$

iff P is an irreflexive relation with for all $a, b, c, d \in X$,

$$(aPb \text{ and } cPd) \Rightarrow (aPd \text{ or } cPb),$$

$$(aPb \text{ and } bPc) \Rightarrow (aPd \text{ or } dPc).$$

In recent papers concerned with decision theory, there appeared a variant of the semiorder model in which two thresholds respectively determine weak and strong preferences. Roy and Hugonnard [10] encode in this way various criteria underlying the forecast analysis of line extensions for the Paris metro, and Vincke [15] shows the adequacy of the model in a case study of a projects comparison. After further advocating their model, Roy and Vincke [11] raise the problem of characterizing pairs of relations amenable to their description. Motivated by other applications whose report shall not be repeated here, Cozzens and Roberts [1] also consider, and solve, the same problem, but leave open the following general question: given m relations $P_1, P_2, \cdots, P_m$ on the finite set X, when do there exist a real-valued mapping f on X and nonnegative real numbers $\sigma_1, \sigma_2, \cdots, \sigma_m$ such that for all $j \in \{1, 2, \cdots, m\}$ and $x, y \in X$

$$xP_j y \Leftrightarrow f(x) > f(y) + \sigma_j. \tag{1}$$

Such families of relations appear in psychological measurement: each P_j captures one level of the preferences an individual expresses among a set X of objects. They are also met in the theory of "probabilistic consistency" (see e.g. Roberts [6], [7] or the references

* Received by the editors December 10, 1984; accepted for publication (in revised form) March 31, 1986. This work was supported by National Science Foundation grant IST84-18860 to J.-Cl. Falmagne at New York University, New York, New York 10012.

† Université Libre de Bruxelles, C.P. 216, Bd du Triomphe, 1050 Bruxelles, Belgium.

therein). Given the frequency $p(a, b)$ that object a is chosen over object b by some given subject, one forms a family P_λ of binary relations by setting for any real number λ

$$aP_\lambda b \quad \text{iff}\, p(a,b) > \lambda.$$

The subject's consistency is then defined in various senses by conditions on the family of relations P_λ. It turns out that the P_λ are usually assumed to be semiorders, and that the existence of a representation as in (1) is a very strong form of consistency.

Now when does a representation as in (1) exist? For $m = 1$, the answer was given in Proposition 1. In case $m = 2$, Cozzens and Roberts [1], relying on a technique due to Scott [13], were able to formulate a rather involved criterion that they could not extend to other values of m. We will provide here a general result for all m, from which easily follows their solution. Our proof uses the so-called potential theory of graphs, and a bit of convex geometry. The relevance of potential theory to semiorders was illustrated by Roy and Vincke [12] in establishing for $m = 2$ a result similar to, but weaker than, the one in [1]. Notice that, independently of us, Roubens and Vincke [9] obtained the Cozzens and Roberts [1] result ($m = 2$).

After having established the general case, we will apply the technique to a still wider setting. The thresholds in the representation can be taken either as constant or depending on the object. Besides semiorders, we thus work in § 5 with interval orders, whose representation theory is due to Fishburn [4].

2. Multiple graphs. Our basic tool is a classical result on graphs that we restate for the reader's convenience. Here a *weighted multiple graph* $G = (V, E, w)$ will be a finite set V of *vertices*, a family E of (ordered) pairs of vertices called *edges* (with repetitions of the same pair allowed), and a *weight mapping* w from E to the reals. By a *cycle* of G we mean any finite sequence of edges having the form $x_1x_2, x_2x_3, \cdots, x_{t-1}x_t, x_tx_1$. The *weight* of a cycle is the sum of the weights of its edges.

PROPOSITION 2. *For $G = (V, E, w)$ as above, the following two conditions are equivalent*:

(i) *there exists a real-valued mapping f on X such that for all $xy \in E$,*

$$f(x) \geqq f(y) + w(xy);$$

(ii) *no cycle of G has a strictly positive weight.*

Clearly in condition (ii) one can replace "cycle" by "simple cycle" (in the sense that no vertex is met more than once by the cycle).

3. Representations of multiple semiorders. Let $\mathcal{P} = (P_1, P_2, \cdots, P_m)$ be an m-tuple of binary relations on the same finite set X. A *constant threshold representation* for $\mathcal{P}$ consists in a real-valued mapping f on X and real numbers $\sigma_1, \sigma_2, \cdots, \sigma_m$ such that the following condition holds for all $j \in \{1, 2, \cdots, m\}$ and $x, y \in X$:

$$xP_j y \Leftrightarrow f(x) > f(y) + \sigma_j.$$

We shall denote this representation by $(f, \sigma_1, \sigma_2, \cdots, \sigma_m)$. When such a mapping f exists, we call the m-tuple of real numbers $(\sigma_1, \sigma_2, \cdots, \sigma_m)$ a *constant threshold vector* for $\mathcal{P}$.

When a representation exists, it is clear that any P_j must be irreflexive (in case $\sigma_j \geqq 0$) or reflexive (in case $\sigma_j < 0$).

For a binary relation R, we write R' for its dual, that is, the converse of its inverse. Given a family $\mathcal{P}$ as above, we call *cycle from* $\mathcal{P}$ any sequence of pairs of the form x_1x_2, $x_2x_3, \cdots, x_{t-1}x_t, x_tx_1$, all taken in $P_1 \cup P'_1 \cup P_2 \cup P'_2 \cup \cdots \cup P_m \cup P'_m$. We associate to this cycle the numbers p_j and p'_j of pairs that were chosen in P_j and P'_j, respectively, and set $q_j = p'_j - p_j$. It will be assumed in the definition of a cycle from $\mathcal{P}$ that $p_j > 0$ for at least one j.

PROPOSITION 3. *The m-tuple* $(\sigma_1, \sigma_2, \cdots, \sigma_m)$ *of real numbers is a constant threshold vector for* $\mathcal{P}$ *iff for each cycle from* $\mathcal{P}$ *the following inequality holds*:

$$q_1\sigma_1 + q_2\sigma_2 + \cdots + q_m\sigma_m > 0.$$

Proof. Assume $(f, \sigma_1, \sigma_2, \cdots, \sigma_m)$ is a representation for $\mathcal{P}$. Since there are only a finite number of values $f(x) - f(y) - \sigma_j$, we can find $\epsilon > 0$ which is a lower bound for all those strictly positive values. We then have

$$xP_j y \Rightarrow f(x) \geqq f(y) + \sigma_j + \epsilon,$$

and also

$$xP'_j y \Rightarrow f(x) \geqq f(y) - \sigma_j.$$

Now define a weighted multiple graph $G = (V, E, w)$ by taking $V = X$ and introducing an edge xy of weight

$$\sigma_j + \epsilon \quad \text{whenever } xP_j y,$$

$$-\sigma_j \quad \text{whenever } xP'_j y.$$

Proposition 2 implies that any cycle of G has nonpositive weight, that is,

$$\sum_{j=1}^{m} p_j(\sigma_j + \epsilon) + \sum_{j=1}^{m} p'_j(-\sigma_j) \leqq 0,$$

or

$$\sum_{j=1}^{m} (p'_j - p_j)\sigma_j - \epsilon \sum_{j=1}^{m} p_j \geqq 0,$$

which implies

$$\sum_{j=1}^{m} q_j\sigma_j > 0.$$

Conversely, since there are only a finite number of cycles from $\mathcal{P}$ without repeated pair, we can find $\epsilon > 0$ such that for any cycle

$$\sum_{j=1}^{m} q_j\sigma_j - \epsilon \sum_{j=1}^{m} p_j \geqq 0.$$

Relying on the same graph as above, and taking the arguments in reverse order, one derives the existence of a mapping f such that $(f, \sigma_1, \sigma_2, \cdots, \sigma_m)$ is a representation for $\mathcal{P}$.

COROLLARY. *There exists a constant threshold vector for* $\mathcal{P}$ *iff there exists such a vector with integer components.*

Let us consider the particular case of two relations P_1 and P_2. Assuming P_2 is irreflexive, a representation (f, σ_1, σ_2) exists for $\mathcal{P} = (P_1, P_2)$ iff such a representation exists with $\sigma_2 > 0$ (this will be explained in the proof of Proposition 4 below). Hence we see that (σ_1, σ_2) is a threshold vector for $\mathcal{P}$ iff for each cycle from $\mathcal{P}$

$$(p'_1 - p_1)\sigma_1 + (p'_2 - p_2)\sigma_2 > 0,$$

which amounts to

$$p'_2 - p_2 > 0 \quad \text{when } p'_1 = p_1,$$

$$\frac{-(p'_2 - p_2)}{p'_1 - p_1} < \frac{\sigma_1}{\sigma_2} \quad \text{when } p'_1 > p_1,$$

$$\frac{\sigma_1}{\sigma_2} < \frac{-(p'_2 - p_2)}{p'_1 - p_1} \quad \text{when } p'_1 < p_1.$$

We deduce that a representation using σ_1, σ_2 exists iff the following four conditions are fulfilled:

(i) no cycle from P_1 uses the same number of pairs from P_1 and P'_1;

(ii) no cycle from P_2 uses at least as many pairs from P_2 than from P'_2;

(iii) no cycle from $\mathscr{P} = (P_1, P_2)$ is balanced, that is, can be obtained by choosing the same number of pairs in P_1 and P'_1, and the same number of pairs in P_2 and P'_2;

(iv) for any two cycles C and $\tilde{C}$ from $\mathscr{P} = (P_1, P_2)$, one has in case $p'_1 > p_1$ and $\tilde{p}'_1 < \tilde{p}_1$:

$$\frac{p_2 - p'_2}{p'_1 - p_1} < \frac{\sigma_1}{\sigma_2} < \frac{\tilde{p}'_2 - \tilde{p}_2}{\tilde{p}_1 - \tilde{p}'_1}.$$

The first two conditions when both P_1 and P_2 are irreflexive mean that these relations are semiorders (cf. Proposition 1), while the last two amount to conditions used by Cozzens and Roberts [1]. Hence, Theorems 6 and 7 of these authors are included in Proposition 3.

4. Characterizations of multiple semiorders. The proposition in the last section does not offer a criterion for the *existence* of a representation for $\mathscr{P} = (P_1, P_2, \cdots, P_m)$. We shall formulate such a criterion using the following concepts. A *k-cyclone* from $\mathscr{P}$ will be any nonempty union of at most k cycles from $\mathscr{P}$. Thus a k-cyclone is obtained by taking pairs in $P_1, P'_1, P_2, P'_2, \cdots, P_m$ and/or P'_m that altogether can be partitioned into k cycles. When for each $j = 1, 2, \cdots, m$, the same number (possibly zero) of pairs is taken in P_j and P'_j, we say that the cyclone is *balanced.*

PROPOSITION 4. *Assume $m \geqq 2$. There is a constant threshold representation for $\mathscr{P} = (P_1, P_2, \cdots, P_m)$ iff no m-cyclone from $\mathscr{P}$ is balanced.*

Proof. First assume that $\mathscr{P}$ admits a representation $(f, \sigma_1, \sigma_2, \cdots, \sigma_m)$. Then

$$xP_j y \Rightarrow f(x) - f(y) > \sigma_j, \qquad xP'_j y \Rightarrow f(x) - f(y) \geqq -\sigma_j.$$

If we consider any k-cyclone, write those implications for all its pairs and sum the resulting inequalities, we get

$$(2) \qquad 0 \geqq \sum_{j=1}^{m} p_j\sigma_j + \sum_{j=1}^{m} p'_j(-\sigma_j) = \sum_{j=1}^{m} (p_j - p'_j)\sigma_j$$

where p_j and p'_j, respectively, denote the number of pairs taken in P_j and P'_j when forming the cyclone. Moreover inequality (2) is strict when at least one pair from some P_j is used. This clearly implies $p_j \neq p'_j$ for at least one j.

Conversely, assume that no m-cyclone from $\mathscr{P}$ is balanced. First notice that any P_j is either reflexive or irreflexive (otherwise we construct a balanced 2-cyclone by taking one loop in P_j and another one in P'_j). Since there are only a finite number of values $f(x) - f(y)$, we can always look for a nonzero threshold σ_1, and by an appropriate change of scale, even assume $\sigma_1 = \pm 1$. We treat the case P_1 is irreflexive, setting $\sigma_1 = +1$. (The other case is similar.)

By Proposition 3, we have to show the existence of a common solution to all the inequalities

$$q_2\sigma_2 + q_3\sigma_3 + \cdots + q_m\sigma_m > -q_1$$

associated to simple cycles from $\mathscr{P}$ (considering simple cycles, as in the remark following Proposition 2, leaves us with a finite number of inequalities). Applying the Helly theorem (see e.g. [8]) to the convex sets defined by those inequalities in the euclidean space of all $(m - 1)$-tuples $(\sigma_2, \sigma_3, \cdots, \sigma_m)$, it is sufficient to show that any m of those inequalities have a common solution, say

$$(3) \qquad q_{i2}\sigma_2+q_{i3}\sigma_3+\cdots+q_{im}\sigma_m>-q_{i1}$$

with $i = 1, 2, \cdots, m$. By a classical result (see e.g. [8, Thm. 22.2]), this is equivalent to the following assertion: given real numbers $\lambda_1, \lambda_2, \cdots, \lambda_m$ with $\lambda_i \geqq 0$ for each i, and $\lambda_i > 0$ for at least one i,

$$(4) \qquad \sum_{i=1}^{m} \lambda_i q_{ij}=0 \quad \text{for } j=2,3, \cdots, m,$$

implies

$$(5) \qquad \sum_{i=1}^{m} \lambda_i q_{i1}>0.$$

Since $q_{ij} = p'_{ij} - p_{ij}$ is an integer, we have only to check the assertion for rational numbers λ_i (because the real tuples $(\lambda_1, \lambda_2, \cdots, \lambda_m)$ satisfying $\lambda_i \geqq 0$ and (4) form a polyhedral convex set with rational extreme points and directions). From the positive homogeneity of (4) and (5) it then follows that we need only consider λ_i's which are natural numbers.

Now any of the equation in (3), for fixed i, comes from a cycle C_i from $\mathscr{P}$. Assuming λ_i is a natural number, we consider the cycle $\tilde{C}_i$ obtained by traversing λ_i times the cycle C_i, and then the union U of those $\tilde{C}_i$. The resulting m-cyclone U uses

$$\sum_{i=1}^{m} \lambda_i p'_{ij} \quad \text{pairs from } P'_j,$$

and

$$\sum_{i=1}^{m} \lambda_i p_{ij} \quad \text{pairs from } P_j.$$

Formula (4) tells us that for $j = 2, 3, \cdots, m$, these two quantities are equal. Since by assumption U is not balanced, we deduce

$$\sum_{i=1}^{m} \lambda_i q_{i1} \neq 0.$$

In order to establish (5), and thus complete the proof, it remains to show that (4) together with

$$(6) \qquad \sum_{i=1}^{m} \lambda_i q_{i1}<0$$

lead to a contradiction. If (6) were true, we would of course have one of the q_{i1} strictly negative (because all λ_i are nonnegative), thus the m-cyclone U would use at least one pair xy from P_1. Now we form a new m-cyclone $\tilde{U}$ by adding

$$\sum_{i=1}^{m} \lambda_i q_{i1}$$

times the pair xx to U (recall that P_1 is taken as irreflexive, thus $xx \in P'_1$). Then $\tilde{U}$ is a balanced m-cyclone, in contradiction with our present assumption.

Notice that Proposition 4 does not directly extend to the case $m = 1$, as exemplified by $X = \{a, b\}$ with $P_1 = \{aa, ab\}$. Nevertheless, it remains true for $m = 1$ under the additional assumption that P_1 be irreflexive or reflexive. The proof prompts a few simple remarks. First, one needs only to ask that m-cyclones of a certain kind be nonbalanced (more precisely, those which are the union of at most m repetitions of simple cycles together with loops). Moreover, this requirement implies that no k-cyclone for any k is balanced.

5. Not necessarily constant thresholds. Fishburn [4] introduced interval orders in connection with representations involving a threshold that depends on the object. More precisely interval orders are the binary relations S on a finite set X that are characterized by the following proposition [4].

PROPOSITION 5. *There exist real valued mappings f, ρ on X such that for all x, $y \in X$,*

$$\rho(x) \geqq 0$$

and

$$xSy \Leftrightarrow f(x) > f(y) + \rho(y)$$

iff S is an irreflexive relation with for all a, b, c, $d \in X$,

$$(aSb \text{ and } cSd) \Rightarrow (aSd \text{ or } cSb).$$

We want now to establish representation theorems for families of relations involving both semiorders and interval orders. Suppose thus that we are given $m + n$ relations P_1, $P_2, \cdots, P_m$, S_1, $S_2, \cdots, S_n$ on the same finite set X, writing for short $\mathscr{P}_{m,n} = (P_1, P_2, \cdots, P_m, S_1, S_2, \cdots, S_n)$. A *threshold representation* for $\mathscr{P}_{m,n}$ consists in real-valued mappings f, $\rho_1, \rho_2, \cdots, \rho_n$ on X and real number $\sigma_1, \sigma_2, \cdots, \sigma_m$ such that the two following equivalences hold for all $x, y \in X$, $j \in \{1, 2, \cdots, m\}$, $l \in \{1, 2, \cdots, n\}$:

$$xP_j y \Leftrightarrow f(x) > f(y) + \sigma_j, \tag{7}$$

$$xS_l y \Leftrightarrow f(x) > f(y) + \rho_l(y). \tag{8}$$

We will need direct generalizations of notions introduced in §§ 3 and 4. A *cycle from* $\mathscr{P}_{m,n}$ is any sequence of pairs of the form x_1x_2, $x_2x_3, \cdots, x_{t-1}x_t$, x_tx_1, taken in

$$P_1 \cup P'_1 \cup P_2 \cup P'_2 \cup \cdots \cup P_m \cup P'_m \cup S_1S'_1 \cup S_2S'_2 \cup \cdots \cup S_nS'_n$$

(where juxtaposition represents relative product), but not all in $P'_1 \cup P'_2 \cup \cdots \cup P'_m$. Then a *k-cyclone* is any nonempty union of at most k cycles, or when $k = 0$, it is taken as being any cycle. It is *simple* if it never meets twice the same element from X. For a given k-cyclone, we denote by p_j, p'_j and s^*_l, respectively, the number of its pairs that were chosen in P_j, P'_j and $S_lS'_l$, and we set also $q_j = p'_j - p_j$. The k-cyclone is *balanced* when $p_j = p'_j$ for $j = 1, 2, \cdots, m$; it is *pure* when $p_j = p'_j = 0$ for $j = 1, 2, \cdots, m$. The qualifiers simple, balanced and pure also apply to cycles.

PROPOSITION 6. *Fix real numbers $\sigma_1, \sigma_2, \cdots, \sigma_m$. There exists a threshold representation for $\mathscr{P}_{m,n}$ involving those numbers iff the following inequality holds for each cycle from $\mathscr{P}_{m,n}$:*

$$q_1\sigma_1 + q_2\sigma_2 + \cdots + q_m\sigma_m > 0,$$

and in particular no pure cycle can be formed.

Proof. First note that the basic equivalences (7) and (8) defining a representation are equivalent to the following four implications

$$xP_j y \Rightarrow f(x) - f(y) > \sigma_j, \tag{9}$$

$$xP'_j y \Rightarrow f(x) - f(y) \geqq -\sigma_j, \tag{10}$$

$$xS_l z \Rightarrow f(x) - f(z) > \rho_l(z),$$

$$zS'_l y \Rightarrow f(z) - f(y) \geqq -\rho_l(z).$$

The last two imply

$$(11) \qquad xS_lS'_l y \Rightarrow f(x) - f(y) > 0.$$

Assuming that a representation exists, consider a cycle with its associated quantities p_j, p'_j and s_l^*. If we write the corresponding implications (9), (10) and (11) for all of its pairs, and sum up the right-hand sides, we obtain

$$0 > \sum_{j=1}^{m} (p_j - p'_j)\sigma_j$$

from which follows the thesis.

Conversely, suppose that we have the inequality in Proposition 6 for each cycle from $\mathscr{P}_{m,n}$, or equivalently, for each of the simple cycles. Since the last ones are in finite number, there exists $\epsilon > 0$ such that for each simple cycle

$$(12) \qquad \sum_{j=1}^{m} q_j\sigma_j \geqq \epsilon\left(\sum_{j=1}^{m} p_j + \sum_{l=1}^{n} s_l^*\right).$$

As in the proof of Proposition 3, we now construct a weighted multiple graph G with vertex set X. Define an edge xy with weight

$$\sigma_j + \epsilon \quad \text{whenever} \quad xP_j y,$$

$$-\sigma_j \quad \text{whenever} \quad xP'_j y,$$

$$\epsilon \quad \text{whenever} \quad xS_lS'_l y.$$

Inequality (12), rewritten as

$$0 \geqq \sum_{j=1}^{m} p_j(\sigma_j + \epsilon) + \sum_{j=1}^{m} p'_j(-\sigma_j) + \sum_{l=1}^{n} s_l^*\epsilon,$$

ensures us that the graph G has no simple cycle with strictly positive sum. By Proposition 2, there exists a real-valued mapping f on X such that

$$xP_j y \Rightarrow f(x) \geqq f(y) + \sigma_j + \epsilon,$$

$$xP'_j y \Rightarrow f(x) \geqq f(y) - \sigma_j,$$

$$(13) \qquad xS_lS'_l y \Rightarrow f(x) \geqq f(y) + \epsilon.$$

From the first two implications follows inequality (7) for a representation. It thus remains to define real-valued mappings ρ_l on X that satisfy (8). We first rewrite (13) as

$$(14) \qquad \text{if } xS_lz \text{ and not } yS_lz, \text{ then } f(x) \geqq f(y) + \epsilon.$$

Now define

$$g_l(z) = \max\{f(y) | \text{not } yS_lz\},$$

agreeing that the maximum of the empty set is some real number less than all values $f(x)$, and then set

$$\rho_l(z) = g_l(z) - f(z).$$

We surely have

$$(\text{not } xS_lz) \Rightarrow f(x) \leqq g_l(z),$$

and also from (14)

$$xS_lz \Rightarrow f(x) > g_l(z).$$

The last two implications amount to (8). This ends the proof.

For families $\mathcal{P}_{0,n}$ consisting only of interval orders, a more direct, purely relational approach leads to the particularization of Proposition 6 (that is, one precludes the existence of alternating cycles in the following sense: pairs of the cycle would be alternatively taken in S_l and the corresponding S'_l). For this and many related results, we refer the reader to Doignon, Monjardet, Roubens and Vincke [3], or Doignon [2].

Finally, a criterion for the existence of a representation can be derived from Proposition 6, exactly as Proposition 4 was derived from Proposition 3. The requirement $m \neq 1$ can be dispensed with as soon as the relation P_1 is assumed to be either reflexive or irreflexive.

PROPOSITION 7. *Assume $m \neq 1$. The family $\mathcal{P}_{m,n}$ admits a threshold representation iff no m-cyclone from $\mathcal{P}_{m,n}$ is balanced.*

REFERENCES

[1] M. B. COZZENS AND F. S. ROBERTS, *Double semiorders and double indifference graphs*, this Journal, 3 (1982), pp. 566–583.

[2] J.-P. DOIGNON, *Generalizations of interval orders*, in Trends in Mathematical Psychology, E. Degreef and J. Van Buggenhaut, eds., North-Holland, Amsterdam, 1984, pp. 209–217.

[3] J.-P. DOIGNON, B. MONJARDET, M. ROUBENS AND PH. VINCKE, *Biorder families, valued relations and preference modelling*, J. Math. Psychol., 30 (1986), to appear.

[4] P. C. FISHBURN, *Intransitive indifference with unequal indifference intervals*, J. Math. Psychol., 7 (1970), pp. 144–149.

[5] R. D. LUCE, *Semiorders and a theory of utility discrimination*, Econometrica, 24 (1956), pp. 178–191.

[6] F. S. ROBERTS, *Homogeneous families of semiorders and the theory of probabilistic consistency*, J. Math. Psychol., 8 (1971), pp. 248-263.

[7] F. S. ROBERTS, *Measurement Theory, with Applications to Decision Making, Utility and the Social Sciences*, Addison-Wesley, Reading, MA, 1979.

[8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[9] M. ROUBENS AND PH. VINCKE, *Preference Modelling*, Lecture Notes in Economics and Mathematical Systems 250, Springer-Verlag, Berlin, 1985.

[10] B. ROY AND J.-CHR. HUGONNARD, *Ranking of suburban line extensions projects on the Paris metro system by a new multicriteria method*, Transportation Res., 16A (1982), pp. 301–312.

[11] B. ROY AND PH. VINCKE, *Relational systems of preference with one or more pseudo-criteria: some new concepts and results*, Management Sci., 30 (1984), pp. 1323–1335.

[12] B. ROY AND PH. VINCKE, *Pseudo-orders: definition, properties and numerical representation*, Math. Social Sci., 14 (1987), to appear.

[13] D. SCOTT, *Measurement models and linear inequalities*, J. Math. Psychol., 1 (1964), pp. 233–247.

[14] D. SCOTT AND P. SUPPES, *Foundational aspects of theories of measurement*, J. Symbolic Logic, 23 (1958), pp. 113–128.

[15] PH. VINCKE, *Preference modelling: a survey and an experiment*, Operational Research '81, J.-P. Brans, ed., North-Holland, Amsterdam, 1981, pp. 341–354.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

006

ON THE NUMBER OF VERTICES OF RANDOM POLYHEDRA WITH A GIVEN NUMBER OF FACETS*

CHRISTIAN BUCHTA†

Abstract. The set of points $\mathbf{x} = (x_1, \cdots, x_n)$ satisfying the linear inequalities $\sum_{\nu=1}^{n} a_{\mu\nu} x_\nu \leqq 1$ $(\mu = 1, \cdots, m)$ is a convex polyhedron. If the m points $\mathbf{a}_\mu = (a_{\mu 1}, \cdots, a_{\mu n})$ are chosen independently and uniformly from the unit sphere in n-space, the number V_{mn} of vertices of the polyhedron is a random variable. We give an asymptotic expansion of the expected value EV_{mn} as $m \rightarrow \infty$ and an explicit formula for EV_{mn} for any m and n.

Key words. random polyhedra

AMS(MOS) subject classifications. Primary 52A22; secondary 60D05

1. Introduction. Two famous results, due, respectively, to McMullen [7] and Barnette [1], state that the number of vertices of a simple bounded polyhedron with exactly m facets is at most

$$\binom{m-\left[\frac{n+1}{2}\right]}{m-n}+\binom{m-\left[\frac{n+2}{2}\right]}{m-n}$$

and at least

$$(m-n)(n-1)+2.$$

For unbounded polyhedra, the lower bound is considerably smaller, being $m - n + 1$. Thus, the upper bound is of order const$(n)m^{[n/2]}$, whereas the lower bound is of order const$(n)m$ as $m \rightarrow \infty$. The gap between the bounds suggests to ask for the number of vertices in the "average" case.

We shall investigate this question for a particular class of polyhedra. We consider the set of points $\mathbf{x} = (x_1, \cdots, x_n)$ satisfying the linear inequalities

$$\sum_{\nu=1}^{n} a_{\mu\nu}x_\nu \leqq 1 \quad \text{where } \sum_{\nu=1}^{n} a_{\mu\nu}^2 = 1 \quad (\mu = 1, \cdots, m).$$

This set is a convex (not necessarily bounded) polyhedron. The condition $\sum_{\nu=1}^{n} a_{\mu\nu}^2 = 1$ implies that the hyperplanes $\sum_{\nu=1}^{n} a_{\mu\nu} x_\nu = 1$ are tangent to the unit sphere. If every $n + 1$ of these m hyperplanes have empty intersection, it follows that the polyhedron has exactly m facets (as no hyperplane is identical to another one) and is simple.

In order to derive the "average" number of vertices of such a polyhedron, we suppose that the points $\mathbf{a}_\mu = (a_{\mu 1}, \cdots, a_{\mu n})$ $(\mu = 1, \cdots, m)$ are chosen independently and uniformly at random from the unit sphere in n-space. Consequently, the number of vertices of the polyhedron $\sum_{\nu=1}^{n} a_{\mu\nu} x_\nu \leqq 1$ $(\mu = 1, \cdots, m)$ is a random variable V_{mn}. Note that, with

* Received by the editors March 11, 1985; accepted for publication (in revised form) March 19, 1986.

† Mathematisches Institut der Universität, Albertstrasse 23 b, D-7800 Freiburg im Breisgau, Federal Republic of Germany. Permanent address: Institut für Analysis, Technische Mathematik und Versicherungsmathematik, Technische Universität, Wiedner Hauptstraße 8-10, A-1040 Wien, Austria.

probability one, every $n + 1$ of the m hyperplanes $\sum_{\nu=1}^{n} a_{\mu\nu}\, x_\nu = 1$ have empty intersection.

Kelly and Tolle [6] investigated the expected value EV_{mn} of V_{mn}: They derived an integral expression for EV_{mn} and asymptotic bounds of the form

$$\alpha^n n^{(n-6)/2} m \leqq EV_{mn} \leqq \beta^n n^{(n-5)/2} m,$$

where the constants α and β are independent of m and n. Moreover, they gave some tables of EV_{mn} computed numerically by means of Gaussian quadrature routine.

In §2 we prove that

$$EV_{mn} = c_0^{(n)} m + c_1^{(n)} m^{1-2/(n-1)} + c_2^{(n)} m^{1-4/(n-1)}$$
$$+ \cdots + c_{[n/2]-1}^{(n)} m^{1-2([n/2]-1)/(n-1)} + O(1) \qquad (m \to \infty).$$

The constants $c_p^{(n)}$ $(p = 0, \cdots, [n/2] - 1)$, which depend on the dimension n of the space, are given explicitly. Especially,

$$c_0^{(n)} = \frac{2^{n-1}\binom{(n-1)^2}{(n-1)^2/2}}{n\binom{n-1}{(n-1)/2}^{n-1}},$$

where, in the case of even n, the binominal coefficients are defined on replacing $n!$ by $\Gamma(n + 1)$. Note that

$$c_0^{(n)} \sim 2^{n/2} \pi^{(n-2)/2} e^{-1/4} n^{(n-5)/2} \qquad (n \to \infty),$$

whence it follows that the upper bound due to Kelly and Tolle gives the exact order of convergence.

In §3 we derive an explicit formula for EV_{mn} for any m and n.

Quite a lot is already known about random polyhedra based on probabilistic models which are different from that considered here. In contrast to the present paper, these polyhedra generally do not have a given number of facets. Important contributions are especially due to Rényi and Sulanke [9], Schmidt [10], Sulanke and Wintgen [12], Prékopa [8] and Schneider [11]. Further references are contained in a recent survey [3]. Prékopa's work is partially extended in [4].

2. An asymptotic expansion of EV_{mn}.

THEOREM 1. *The expected number EV_{mn} of vertices of the polyhedron $\sum_{\nu=1}^{n} a_{\mu\nu} x_\nu \leqq 1$ $(\mu = 1, \cdots, m)$, where the m points $\mathbf{a}_\mu = (a_{\mu 1}, \cdots, a_{\mu n})$ are chosen independently and uniformly at random from the unit sphere in n-space, is given by*

$$EV_{mn} = \sum_{p=0}^{[n/2]-1} c_p^{(n)} m^{1-2p/(n-1)} + O(1) \qquad (m \to \infty).$$

The constants $c_p^{(n)}$ are defined by

$$c_p^{(n)} = 2^{1-2p} \frac{n-1}{n!} \gamma_{(n-1)^2} \gamma_{n-1}^{-n+1-2p/(n-1)} \sum_{i=0}^{[n/2]-1} d_{ip}^{(n)} \Gamma\left(n + i - 1 + \frac{2p}{n-1}\right),$$

where $\gamma_n = 2^{-(n+1)} \Gamma(n + 1)\{\Gamma(n/2) + 1)\}^{-2}$ and where $d_{ip}^{(n)}$ is the coefficient of x^p in the polynomial

$$P_i^{(n)}(x)=\frac{(-1)^i}{i!}\left(\sum_{j=1}^{[n/2]-1}(-1)^j\binom{(n-3)/2}{j}\frac{n-1}{n+2j-1}x^j\right)^i$$
$$\cdot\left(\sum_{k=0}^{[n/2]-1}(-1)^k\binom{(n^2-2n-1)/2}{k}x^k\right).$$

Proof. For the sake of completeness, we first sketch the idea of Kelly and Tolle leading to the integral expression of EV_{mn}. Consider the hyperplanes

$$\sum_{\nu=1}^{n} a_{\mu\nu}x_\nu=1 \qquad (\mu=1,\cdots,n).$$

The n points $\mathbf{a}_\mu=(a_{\mu 1},\cdots,a_{\mu n})$ lie on a hypercircle which divides the unit sphere S_{n-1} into two caps; we denote the surface area of the smaller cap by $\tilde{S}=\tilde{S}(\mathbf{a}_1,\cdots,\mathbf{a}_n)$. The intersection of the n hyperplanes is a vertex of the polyhedron in question if none of the points $\mathbf{a}_\mu=(a_{\mu 1},\cdots,a_{\mu n})$ $(\mu=n+1,\cdots,m)$ lies on the smaller cap. As all points are independently and uniformly distributed, this event occurs with probability

$$\left(1-\frac{\tilde{S}}{\omega_n}\right)^{m-n},$$

where ω_n denotes the surface area of the unit sphere in n-space. As the points $\mathbf{a}_\mu=(a_{\mu 1},\cdots,a_{\mu n})$ $(\mu=1,\cdots,m)$ are identically distributed and as there are $\binom{m}{n}$ possibilities of choosing n points out of m, it follows that

$$EV_{mn}=\binom{m}{n}\int_{S_{n-1}}\cdots\int_{S_{n-1}}\left(1-\frac{\tilde{S}}{\omega_n}\right)^{m-n}\frac{d\omega(\mathbf{a}_1)}{\omega_n}\cdots\frac{d\omega(\mathbf{a}_n)}{\omega_n},$$

where ω is the spherical surface measure. A transformation due to Miles yields

$$EV_{mn}=2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_0^{\pi/2}\left(1-\frac{\omega_{n-1}}{\omega_n}\int_0^r\sin^{n-2}x\,dx\right)^{m-n}\sin^{n^2-2n}r\,dr;$$

for details cf. the paper of Kelly and Tolle. (We use the symbol γ_n instead of Kelly and Tolle's symbol $c(n)$. It is easy to see that $\gamma_n=\{\pi c(n)\}^{-1}$.)

Putting $1-\cos r=s$ and $1-\cos x=y$, we obtain

$$EV_{mn}=2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_0^1(1-K_n(s))^{m-n}(s(2-s))^{(n^2-2n-1)/2}ds,$$

where

$$K_n(s)=\frac{\omega_{n-1}}{\omega_n}\int_0^s(y(2-y))^{(n-3)/2}dy.$$

We divide this integral into

$$I_1=2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_0^{1/2}(1-K_n(s))^{m-n}(s(2-s))^{(n^2-2n-1)/2}ds,$$

$$I_2=2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_{1/2}^1(1-K_n(s))^{m-n}(s(2-s))^{(n^2-2n-1)/2}ds.$$

Obviously,

$$I_2<2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_{1/2}^{1}(1-K_n(\tfrac{1}{2}))^{m-n}(s(2-s))^{(n^2-2n-1)/2}\,ds.$$

As $K_n(s)$ is the ratio of the surface area of a cap of height s to the surface area of the unit sphere, it follows that $0<1-K_n(\frac{1}{2})<1$. Thus, I_2 exponentially tends to zero as m tends to infinity.

To determine the asymptotic behaviour of I_1, we note that

$$\left(1-\frac{y}{2}\right)^{(n-3)/2}=\sum_{j=0}^{[n/2]-1}(-1)^j\binom{(n-3)/2}{j}\left(\frac{y}{2}\right)^j$$
$$+(-1)^{[n/2]}\binom{(n-3)/2}{[n/2]}\left(1-\frac{\theta y}{2}\right)^{(n-3)/2-[n/2]}\left(\frac{y}{2}\right)^{[n/2]},$$

$$\left(1-\frac{s}{2}\right)^{(n^2-2n-1)/2}=\sum_{k=0}^{[n/2]-1}(-1)^k\binom{(n^2-2n-1)/2}{k}\left(\frac{s}{2}\right)^k$$
$$+(-1)^{[n/2]}\binom{(n^2-2n-1)/2}{[n/2]}\left(1-\frac{\xi s}{2}\right)^{(n^2-2n-1)/2-[n/2]}\left(\frac{s}{2}\right)^{[n/2]},$$

where $0<\theta<1$ and $0<\xi<1$. We now replace $m-n$ by m and put

$$t=m\gamma_{n-1}(2s)^{(n-1)/2};$$

taking into consideration that $\gamma_{n-1}=\omega_{n-1}\{(n-1)\omega_n\}^{-1}$, we obtain

$$EV_{mn}=2\binom{m+n}{n}(n-1)\gamma_{(n-1)^2}\int_0^{m\gamma_{n-1}}\frac{1}{m\gamma_{n-1}}\left(\frac{t}{m\gamma_{n-1}}\right)^{n-2}$$
$$\cdot\left(1-\frac{t}{m}\sum_{j=0}^{[n/2]-1}(-1)^j\binom{(n-3)/2}{j}\frac{n-1}{n+2j-1}2^{-2j}\left(\frac{t}{m\gamma_{n-1}}\right)^{2j/(n-1)}\right)^m$$
$$\cdot\left(\sum_{k=0}^{[n/2]-1}(-1)^k\binom{(n^2-2n-1)/2}{k}2^{-2k}\left(\frac{t}{m\gamma_{n-1}}\right)^{2k/(n-1)}\right)dt+O(1)\qquad(m\to\infty).$$

As

$$\left(1-\frac{t}{m}\sum_{j=0}^{[n/2]-1}(-1)^j\binom{(n-3)/2}{j}\frac{n-1}{n+2j-1}2^{-2j}\left(\frac{t}{m\gamma_{n-1}}\right)^{2j/(n-1)}\right)^m$$
$$=\sum_{i=0}^{[n/2]-1}\frac{(-t)^i}{i!}\left(\sum_{j=1}^{[n/2]-1}(-1)^j\binom{(n-3)/2}{j}\frac{n-1}{n+2j-1}2^{-2j}\left(\frac{t}{m\gamma_{n-1}}\right)^{2j/(n-1)}\right)^i$$
$$\cdot\left(1-\frac{t}{m}\right)^m+O\left(\frac{1}{m}\right)\qquad(m\to\infty),$$

it follows that

$$EV_{mn}=2\frac{n-1}{n!}\gamma_{(n-1)^2}\sum_{p=0}^{[n/2]-1}2^{-2p}\gamma_{n-1}^{-n+1-2p/(n-1)}$$
$$\cdot\sum_{i=0}^{[n/2]-1}d_{ip}^{(n)}\int_0^{m\gamma_{n-1}}\left(1-\frac{t}{m}\right)^m t^{n+i-2+2p/(n-1)}\,dt$$
$$\cdot\,m^{1-2p/(n-1)}+O(1)\qquad(m\to\infty).$$

To get rid of the remaining integrals, we use the relation

$$\int_0^{rm}\left(1-\frac{t}{m}\right)^m t^\alpha\,dt=\Gamma(\alpha+1)+O\left(\frac{1}{m}\right)\qquad(m\to\infty),$$

which holds for $0 < r \leqq 1$ and $\alpha > 0$. (For a proof of this relation cf., e.g., [2, Lemma 1].) Elementary calculations yield $\gamma_0 = \frac{1}{2}$, $\gamma_1 = 1/\pi$, $\gamma_{n+1} = n\gamma_{n-1}/(n+1)$, whence $\gamma_n \leqq \frac{1}{2}$ for any n. Thus,

$$\int_0^{m\gamma_{n-1}}\left(1-\frac{t}{m}\right)^m t^{n+i-2+2p/(n-1)}\,dt=\Gamma\left(n+i-1+\frac{2p}{n-1}\right)+O\left(\frac{1}{m}\right)\qquad(m\to\infty),$$

and we obtain the claimed expansion

$$EV_{mn}=2\frac{n-1}{n!}\gamma_{(n-1)^2}\sum_{p=0}^{[n/2]-1}2^{-2p}\gamma_{n-1}^{-n+1-2p/(n-1)}$$

$$\cdot\sum_{i=0}^{[n/2]-1}d_{ip}^{(n)}\Gamma\left(n+i-1+\frac{2p}{n-1}\right)m^{1-2p/(n-1)}+O(1)\qquad(m\to\infty).\qquad\square$$

COROLLARY.

$$EV_{mn}\sim\frac{2^{n-1}\begin{pmatrix}(n-1)^2\\(n-1)^2/2\end{pmatrix}}{n\begin{pmatrix}n-1\\(n-1)/2\end{pmatrix}^{n-1}}m\qquad(m\to\infty).$$

Proof. For any n,

$$d_{i,0}^{(n)}=\begin{cases}1 & \text{for } i=0,\\ 0 & \text{for } i=1,\cdots,[n/2]-1,\end{cases}$$

hence

$$c_0^{(n)}=\frac{2}{n}\gamma_{(n-1)^2}\gamma_{n-1}^{-(n-1)}=\frac{2^{n-1}\begin{pmatrix}(n-1)^2\\(n-1)^2/2\end{pmatrix}}{n\begin{pmatrix}n-1\\(n-1)/2\end{pmatrix}^{n-1}}.\qquad\square$$

To illustrate Theorem 1, we give some numerical values:

$$EV_{m2}=1{,}00m+O(1)\qquad(m\to\infty),$$

$$EV_{m3}=2{,}00m+O(1)\qquad(m\to\infty),$$

$$EV_{m4}=6{,}77m-22{,}90m^{1/3}+O(1)\qquad(m\to\infty),$$

$$EV_{m5}=31{,}78m-142{,}27m^{1/2}+O(1)\qquad(m\to\infty),$$

$$EV_{m6}=186{,}74m-1007{,}64m^{3/5}+1778{,}82m^{1/5}+O(1)\qquad(m\to\infty).$$

3. An explicit formula for EV_{mn}.

THEOREM 2. *The expected number EV_{mn} of vertices of the polyhedron $\sum_{\nu=1}^{n} a_{\mu\nu}x_\nu \leqq 1$ ($\mu = 1, \cdots, m$), where the m points $\mathbf{a}_\mu = (a_{\mu 1}, \cdots, a_{\mu n})$ are chosen independently and uniformly at random from the unit sphere in n-space, is given by*

$$EV_{mn}=\begin{cases} 2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\sum_{k=0}^{m-n}(-1)^k\binom{m-n}{k}\gamma_0^{m-n-k} \\ \quad\cdot\sum_{j=0}^{(n-3)k/2}\langle\gamma_0,\gamma_2,\cdots,\gamma_{n-3}\rangle_j^k J(0,n^2-2n+2j,k) \quad \textit{for odd } n, \\ 2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\sum_{k=0}^{m-n}(-1)^k\binom{m-n}{k}\gamma_1^{m-n-k} \\ \quad\cdot\sum_{j=0}^{(n-4)k/2}\langle\gamma_1,\gamma_3,\cdots,\gamma_{n-3}\rangle_j^k J(m-n-k,n^2-2n+2j+k,k) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{for even } n, \end{cases}$$

where $\gamma_n = 2^{-(n+1)}\Gamma(n+1)\{\Gamma((n/2)+1)\}^{-2}$, $\langle c_0,\cdots,c_r\rangle_j^k$ *denotes the coefficient of* x^j *in the polynomial* $(\sum_{i=0}^r c_i x^i)^k$, *and*

$$J(l,p,q)=\int_{\pi/2}^{\pi} t^l \sin^p t \cos^q t\,dt.$$

Proof. As shown in the proof of Theorem 1,

$$EV_{mn}=2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_0^{\pi/2}\left(1-\frac{\omega_{n-1}}{\omega_n}\int_0^r \sin^{n-2}x\,dx\right)^{m-n}\sin^{n^2-2n}r\,dr.$$

From the relation

$$1-\frac{\omega_{n-1}}{\omega_n}\int_0^r \sin^{n-2}x\,dx=\frac{\omega_{n-1}}{\omega_n}\int_r^\pi \sin^{n-2}x\,dx$$

it follows on putting $t=\pi-r$ and $z=\pi-x$ that

$$EV_{mn}=2\binom{m}{n}(n-1)^2\gamma_{(n-1)^2}\int_{\pi/2}^{\pi}\left(\frac{\omega_{n-1}}{\omega_n}\int_0^t \sin^{n-2}z\,dz\right)^{m-n}\sin^{n^2-2n}t\,dt.$$

TABLE 1

EV_{mn}/m	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
$m=n+1$	0,7500	0,6875	0,6594	0,6418	0,6292
$m=n+2$	0,8750	0,9750	1,1162	1,2647	1,4146
$m=n+3$	0,9375	1,1875	1,5521	1,9846	2,4750
$m=n+4$	0,9688	1,3393	1,9381	2,7331	3,7332
$m=n+5$	0,9844	1,4473	2,2685	3,4666	5,1134
$m=n+6$	0,9922	1,5252	2,5481	4,1628	6,5560
$m=n+7$	0,9961	1,5828	2,7851	4,8132	8,0196
$m=n+8$	0,9980	1,6268	2,9879	5,4171	9,4781
$m=n+9$	0,9990	1,6614	3,1637	5,9778	10,9168
$m=n+10$	0,9995	1,6894	3,3179	6,4996	12,3279
⋮	⋮	⋮	⋮	⋮	⋮
$m\to\infty$	1,0000	2,0000	6,7677	31,7778	186,7380

Using the identity

$$\frac{\omega_{n-1}}{\omega_n}\int_0^t \sin^{n-2} z\,dz = \begin{cases} \gamma_0 - \cos t \sum\limits_{j=0}^{(n-3)/2} \gamma_{2i}\sin^{2i} t & \text{for odd } n, \\ \gamma_1 t - \cos t \sum\limits_{j=0}^{(n-4)/2} \gamma_{2i+1}\sin^{2i+1} t & \text{for even } n, \end{cases}$$

we obtain

$$\left(\frac{\omega_{n-1}}{\omega_n}\int_0^t \sin^{n-2} z\,dz\right)^{m-n}$$

$$= \begin{cases} \sum\limits_{k=0}^{m-n} (-1)^k \binom{m-n}{k} \gamma_0^{m-n-k} \\ \quad \cdot \sum\limits_{j=0}^{(n-3)/2} \langle \gamma_0, \gamma_2, \cdots, \gamma_{n-3}\rangle_j^k \cos^k t \sin^{2j} t & \text{for odd } n, \\ \sum\limits_{k=0}^{m-n} (-1)^k \binom{m-n}{k} \gamma_1^{m-n-k} \\ \quad \cdot \sum\limits_{j=0}^{(n-4)/2} \langle \gamma_1, \gamma_3, \cdots, \gamma_{n-3}\rangle_j^k t^{m-n-k} \cos^k t \sin^{2j+k} t & \text{for even } n, \end{cases}$$

and hence the claimed result. □

Table 1 gives some values of EV_{mn}/m calculated by means of Theorem 2. The ratio of the expected number of vertices to the number of inequalities provides a better insight into the geometrical situation than the mere value EV_{mn}.

4. Concluding remarks. (*a*) Denote by p_{mn} the probability that the considered polyhedron is unbounded. As mentioned by Kelly and Tolle, $p_{m2} = m/2^{m-1}$. More generally, already in 1962, Wendel [13] showed by an elegant argument that

$$p_{mn} = \frac{1}{2^{m-1}} \sum_{k=0}^{n-1} \binom{m-1}{k}.$$

(*b*) The expected value EV_{mn} is closely related to the expected number EF_{mn} of facets of the convex hull of m random points chosen independently and uniformly from a sphere in n-space, which was determined in a recent paper [5, Thm. 3]. Using the above notation,

$$EF_{mn} - EV_{mn} = \binom{m}{n} \int_{S_{n-1}} \cdots \int_{S_{n-1}} \left(\frac{\tilde{S}}{\omega_n}\right)^{m-n} \frac{d\omega(\mathbf{a}_1)}{\omega_n} \cdots \frac{d\omega(\mathbf{a}_n)}{\omega_n},$$

and thus $EV_{mn} < EF_{mn}$. Further, $\tilde{S}/\omega_n \leqq \frac{1}{2}$, hence it follows that $EV_{mn} \sim EF_{mn}$ as m tends to infinity. (Note that the definition of $J(l, p, q)$ in Theorem 2 of the present article is different from the definition of $I(m, p, q)$ in [5, §2].)

REFERENCES

[1] D. W. BARNETTE, *The minimum number of vertices of a simple polytope*, Israel J. Math., 10 (1971), pp. 121–125.

[2] C. BUCHTA, *Stochastische Approximation konvexer Polygone*, Z. Wahrsch. Verw. Gebiete, 67 (1984), pp. 283-304.

[3] ———, *Zufällige Polyeder - Eine Übersicht*, in Zahlentheoretische Analysis, E. Hlawka, ed., Lecture Notes in Mathematics, 1114, Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1985, pp. 1-13.

[4] ———, *On non-negative solutions of random systems of linear inequalities*, Disc. Comput. Geometry, to appear.

[5] C. BUCHTA, J. MÜLLER AND R. F. TICHY, *Stochastical approximation of convex bodies*, Math. Ann., 271 (1985), pp. 225-235.

[6] D. G. KELLY AND J. W. TOLLE, *Expected number of vertices of a random convex polyhedron*, this Journal, 2 (1981), pp. 441-451.

[7] P. MCMULLEN, *The maximum number of faces of a convex polytope*, Mathematika, 17 (1970), pp. 179-184.

[8] A. PRÉKOPA, *On the number of vertices of random convex polyhedra*, Period. Math. Hungar., 2 (1972), pp. 259-282.

[9] A. RÉNYI AND R. SULANKE, *Zufällige konvexe Polygone in einem Ringgebiet*, Z. Wahrsch. Verw. Gebiete, 9 (1968), pp. 146-157.

[10] W. M. SCHMIDT, *Some results in probabilistic geometry*, Z. Wahrsch. Verw. Gebiete, 9 (1968), pp. 158-162.

[11] R. SCHNEIDER, *Random polytopes generated by anisotropic hyperplanes*, Bull. London Math. Soc., 14 (1982), pp. 549-553.

[12] R. SULANKE AND P. WINTGEN, *Zufällige konvexe Polyeder im n-dimensionalen euklidischen Raum*, Period. Math. Hungar., 2 (1972), pp. 215-221.

[13] J. G. WENDEL, *A problem in geometric probability*, Math. Scand., 11 (1962), pp. 109-111.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

007

THE EXISTENCE OF A SUBSQUARE FREE LATIN SQUARE OF SIDE 12*

P. B. GIBBONS† AND E. MENDELSOHN‡

Abstract. A subsquare free Latin square of side 12 is displayed. The computational method for its construction is outlined, and its significance is discussed.

Key words. Latin square, quasigroup, subsquare, backtracking

AMS(MOS) subject classifications. 05B15, 68E99

1. Introduction. In this paper a Latin square of side n will be an $n \times n$ matrix with entries from the set of n integers $\{1, 2, \cdots, n\}$. A quasigroup is an algebra with one binary operation whose multiplication table is a Latin square. Thus we have a distinction between a subsquare and a subquasigroup. A subsquare of side k of a Latin square is a triple R, C, E, $|R| = |C| = |E|$, such that if $R = \{r_1, r_2, \cdots, r_k\}$ is a set of row names, and $C = \{c_1, c_2, \cdots, c_k\}$ is a set of column names, then the entry in the r_ith row and c_jth column ($i, j = 1, 2, \cdots, k$) is always contained in E. A subquasigroup requires in addition that $R = C = E$.

The first major step in the resolution of the problem of the existence of Latin squares with no proper subsquares was made by Heinrich [5] who managed to construct subsquare free Latin squares (SFLS's) of order $n = pq$, where p and q are distinct primes, and $n \neq 6$. Her method was extended and generalized by Andersen and Mendelsohn [2], who showed that such squares exist for all n not of the form $n = 2^a 3^b$. When n *is* of the form $2^a 3^b$ it is known that SFLS's do *not* exist for $n = 4$, 6, and that they *do* exist for $n = 8$ (see [3], [5] and [6]).

There is also a relationship between one-factorizations and SFLS's. From a one-factorization of K_n an idempotent Latin square of side $n - 1$ can be constructed. If the one-factorization is perfect, then the union of any pair of distinct one-factors forms a Hamiltonian cycle in K_n. This means that in the corresponding Latin square, the smallest subsquare containing any pair of distinct elements must be of size $n - 1$. That is, the Latin square is subsquare free. (For details the reader is referred to the survey paper of Mendelsohn and Rosa [8].) Perfect one-factorizations of K_n are known to exist when $n - 1$ is an odd prime, when $n/2$ is a prime, and when $n = 16$, 28, 244, and 344. (See [8] for the relevant references.) Thus, for example, there exist SFLS's of orders 3, 9, 27, 81, and 243, but for no other known orders $n = 3^b \leqq 3^{12}$.

From the above we see that the first unsolved cases for SFLS's are $n = 12$, 16, and 18.

In other related work N. S. Mendelsohn ([9]) showed that for all n there exists a quasigroup of order n with no proper subquasigroup. Kotzig, Lindner and Rosa [4] showed that if $n \neq 2^a$ there is a 2×2 subsquare free Latin square (N2LS). McLeish [7] showed that an N2LS exists for $n = 2^a$, $a \geqq 6$. Kotzig and Turgeon [6] constructed N2LS's

* Received by the editors September 30, 1985; accepted for publication (in revised form) May 27, 1986.

† Department of Computer Science, University of Auckland, Auckland, New Zealand.

‡ Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A4. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant #A7621 and by the Stanley Vineberg Fund during a sabbatical at the Technion, Israel Institute of Technology.

for all even orders $n \neq 0 \pmod 3$, $n \neq 3 \pmod 5$, thus establishing existence for the special cases $n = 2^4, 2^5$. They also presented an N2LS of order 8 due to Regener. Thus N2LS's exist for all $n \neq 2, 4$. Although the N2LS's in these constructions contain larger subsquares, the idea of N_2-completion of a subsquare free Latin rectangle turned out to be a useful heuristic in the search for a SFLS of order 12.

Finally in this section we note that the SFLS construction theorem of Andersen and Mendelsohn [2] states that there is an SFLS of order pm where p is a prime greater than 3 and m is any positive integer. It is hoped that a multiplication by 8 and 9 using Regener's square and the SFLS of order 9 in place of the cyclic square of prime order might someday be found. This would leave only the case $n = 12$ in doubt. This paper removes that doubt.

2. The construction. We begin this section by investigating conditions for a given $m \times n$ $(1 < m \leqq n)$ Latin rectangle L to be subsquare free. In L any pair of rows defines a permutation of the n elements—we shall call this a row permutation. The Latin rectangle property prescribes that there is no row permutation containing fixed elements, i.e. cycles of length 1. In addition, for L to be 2×2 subsquare free (N_2), a necessary and sufficient condition is that there is no row permutation containing a 2-cycle. Unfortunately this does not generalize for L to be $p \times p$ $(2 < p \leqq m)$ subsquare free (N_p). However we can say that a necessary condition for L to be N_p is that there is no row which forms p-cycles with $p - 1$ other rows on a common set of p elements. In attempting to construct an N_p Latin square L we could enforce the stronger condition that no row form p-cycles with more than $p - 2$ other rows (whether on a common set of p elements or not). This is the condition that we exploited in attempting to construct Latin rectangles, and hopefully Latin squares, which were completely subsquare free.

As our aim is to produce an SFLS(12) we have the following extra information:

(a) The largest possible subsquare is of side 6.

(b) If a 6×6 subsquare exists then either a 2×2 or a 3×3 subsquare also exists.

(c) If a 5×5 subsquare exists then either

(i) there are 5 rows each pair of which contains a 5-cycle in its row permutation (and on the same set of cells), or

(ii) there is a 2×2 subsquare.

(d) The existence of a 4×4 subsquare implies the existence of a 2×2 subsquare.

(e) The existence of a 3×3 subsquare implies the existence of a set of 3 rows each pair of which has a 3-cycle in its row permutation (and on the same set of cells).

During the construction we placed restrictions on the types of row permutations that might be formed. In particular we encouraged the use of row permutations containing long cycles, in the hope that this would allow a large number of rows to be constructed which should not contain a subsquare. In our case, $n = 12$, and the cycle types in our choice of decreasing order of desirability are (12), (6, 6), (5, 7), (4, 8), (4, 4, 4), (3, 9), (3, 4, 5), (3, 3, 6), and (3, 3, 3, 3) (remembering of course that 2-cycles are banned completely). The obvious starting point was to attempt to construct a square composed only of 12-cycles. However this was found to be impossible. Our next attempt involved using only the cycle types (12) and (6, 6). If a rectangle could be constructed containing only these cycle types, it would be subsquare free. This suggested the use of a backtracking algorithm to construct the rectangle row by row, allowing each new row to form a (12) or (6, 6) cycle type with each previous row, and making sure that the new row forms a (6, 6) cycle type with at most 4 previous rows.

This approach was generalized to allow other mixes of cycle types. With each permissible cycle type $c = (a_1, a_2, \cdots, a_q)$, $3 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_q$, we allowed each new row to form a cycle type *c* with a maximum of $a_l - 2$ previous rows. (A stronger restriction would have been to allow no new row to form *any* cycle type with the same a_l with more than $a_l - 2$ previous rows. However this was not implemented.) Rows were constructed in increasing lexicographical order. Moreover, for each row, all possible combinations of allowable cycle types with previous rows were considered. The algorithm is summarized by the following Pascal-like pseudocode:

```
{Attempt to construct an $n \times n$ Latin square with specified allowable cycle types and
 limits on the use of each cycle type.}
begin
  {Get set to start search}
  Set first row and column of square to $(1, 2, \cdots, n)$;
  current_row:= 1;
next_row:
  {Square found?}
  if current_row = n
  then begin
       {Search succeeds}
       Output specified pattern Latin square;
       halt;
       end;
  {Advance to next row}
  current_row: = current_row+1;
  Initialise cycle pattern of current_row with previous rows;
try_row:
  {Construct row}
  Attempt to construct next current_row in lexicographical order according to specified
  cycle pattern with previous rows;
  {Successful?}
  if successful
  then go to next_row;
  {Adjust cycle pattern}
  Find next cycle pattern of current_row with previous rows;
  if there is a next pattern
  then begin
       Prepare to start afresh with construction of current_row;
       go to try_row;
       end
  else begin
       {Backtrack to previous row}
       if current_row = 2
         then begin
              {Search fails}
              Output "Search failed—no such Latin Square";
              halt;
             end;
         current_row:= current_row−1;
         go to try_row;
            end;
end.
```

Note that we were actually searching for a square of the form

$$\begin{array}{cccccccc}
1 & 2 & 3 & \cdot & \cdot & \cdot & n \\
2 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
3 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
n & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}$$

This was not unduly restrictive since any completed square could be transformed into such a form by performing appropriate row and column permutations. Also note that we did not give up trying to construct a particular row until all possible cycle patterns with previous rows had been tried.

The algorithm was implemented initially in the language Pascal on a Hewlett Packard 9836 16-bit microcomputer, and later in *C* on a PDP/11. Several cycle patterns were tried, and in most cases the program had great difficulty in constructing more than the first 7 or 8 rows of the square. In some cases a 9 × 12 rectangle was constructed, but in no cases was the program able to proceed any further.

We then modified our approach so that the above algorithm was used only up to row 8, at which stage we were very likely to have a subsquare free Latin rectangle. We then tried to complete each such rectangle to an N_2 Latin square, hoping that the N_2 constraint would be strong enough to preclude the existence of larger subsquares. Using the allowable cycle types (12), (6, 6), and (5, 7) a large number of 8 × 12 rectangles were constructed and tested for N_2 completion. Eventually the following rectangle, which can be N_2 completed, was produced by the program. It turned out that its completion is completely subsquare free, as we shall show.

$$\begin{array}{cccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
2 & 3 & 4 & 5 & 6 & 1 & 8 & 9 & 10 & 11 & 12 & 7 \\
3 & 1 & 5 & 2 & 7 & 8 & 4 & 10 & 6 & 12 & 9 & 11 \\
4 & 5 & 6 & 7 & 1 & 9 & 11 & 12 & 8 & 3 & 2 & 10 \\
5 & 6 & 2 & 8 & 10 & 7 & 9 & 11 & 12 & 4 & 1 & 3 \\
6 & 12 & 8 & 1 & 3 & 10 & 2 & 7 & 11 & 9 & 4 & 5 \\
7 & 8 & 1 & 10 & 12 & 11 & 5 & 4 & 2 & 6 & 3 & 9 \\
8 & 9 & 11 & 3 & 4 & 12 & 10 & 6 & 5 & 1 & 7 & 2
\end{array}$$

If we denote the allowable cycle types as follows:

1:	(6,6)
2:	(12)
3:	(5,7)

then the row permutation cycle type structure of the above rectangle is denoted by the following matrix C:

$$\begin{array}{l|cccccccc}
1: & - & 1 & 1 & 1 & 1 & 1 & 2 & 2 \\
2: & 1 & - & 1 & 1 & 1 & 1 & 2 & 2 \\
3: & 1 & 1 & - & 3 & 3 & 3 & 2 & 2 \\
4: & 1 & 1 & 3 & - & 3 & 3 & 2 & 2 \\
5: & 1 & 1 & 3 & 3 & - & 3 & 2 & 2 \\
6: & 1 & 1 & 3 & 3 & 3 & - & 2 & 2 \\
7: & 2 & 2 & 2 & 2 & 2 & 2 & - & 3 \\
8: & 2 & 2 & 2 & 2 & 2 & 2 & 3 & -
\end{array}$$

where C_{ij} is the cycle type of the permutation formed by rows i and j. From C it is immediately clear that the rectangle contains no subsquares of order 5, since no row in C contains more than three 3's. Furthermore, although row 1 forms 6-cycles with rows 2 through 6, these rows cannot contain a 6 × 6 subsquare since there are no 2 × 2 or 3 × 3 subsquares. Thus the rectangle is completely subsquare free. Furthermore, this rectangle was able to be completed to the following N_2 Latin square:

1	2	3	4	5	6	7	8	9	10	11	12
2	3	4	5	6	1	8	9	10	11	12	7
3	1	5	2	7	8	4	10	6	12	9	11
4	5	6	7	1	9	11	12	8	3	2	10
5	6	2	8	10	7	9	11	12	4	1	3
6	12	8	1	3	10	2	7	11	9	4	5
7	8	1	10	12	11	5	4	2	6	3	9
8	9	11	3	4	12	10	6	5	1	7	2
9	11	7	12	2	5	1	3	4	8	10	6
10	7	12	11	9	4	6	1	3	2	5	8
11	4	10	9	8	3	12	2	7	5	6	1
12	10	9	6	11	2	3	5	1	7	8	4

with the following row permutation cycle structure:

1:	-	1	1	1	1	1	2	2	2	2	2	5
2:	1	-	1	1	1	1	2	2	2	2	4	1
3:	1	1	-	3	3	3	2	2	4	2	2	3
4:	1	1	3	-	3	3	2	2	2	2	2	6
5:	1	1	3	3	-	3	2	2	7	2	2	3
6:	1	1	3	3	3	-	2	2	2	2	7	3
7:	2	2	2	2	2	2	-	3	3	6	5	2
8:	2	2	2	2	2	2	3	-	1	3	3	2
9:	2	2	4	2	7	2	3	1	-	5	6	2
10:	2	2	2	2	2	2	6	3	5	-	5	2
11:	2	4	2	2	2	7	5	3	6	5	-	2
12:	5	1	3	6	3	3	2	2	2	2	2	-

where the cycle types are denoted as follows:

1:	(6,6)
2:	(12)
3:	(5,7)
4:	(3,4,5)
5:	(3,9)
6:	(4,8)
7:	(3,3,6)

It is not immediately obvious that the above Latin square contains no subsquares. For example, notice that row 10 forms 3-cycles with rows 9 and 11. However row 9 does not form a 3-cycle with row 11, so that rows 9, 10 and 11 cannot contain a 3×3 subsquare. Similarly, a small number of other cases can be checked by hand to verify that in fact the square is completely subsquare free.

To be quite certain, we constructed a series of graphs G_i, $i = 2, 3, 5$, in which the vertices of G_i represent the rows of the above square, and in which two vertices h and k are adjacent if and only if rows h and k form a permutation containing an i-cycle. A clique analysis was performed on each such graph G_i to determine that it contained no i-cliques, and therefore that the square contained no $i \times i$ subsquares.

3. Conclusions. The search for subsquare free Latin squares raises questions of a different nature from some of the standard ones. The usual Latin square questions are of the form "Can a set of cells be completed to form an $n \times n$ Latin square?", or "Can a quasigroup in a given variety be completed to a quasigroup within the variety?". (The reader is referred to [1] for details.) Our sort of restricted completion is much more difficult to analyse. In fact we would propose the following two open questions:

(a) What is the optimal time for an algorithm that will check whether a given Latin square is subsquare free? The complexity of such an algorithm must be at least $O(n^3)$, as that is the time needed to check that it is N_2. (It is possible to construct Latin squares of prime order with only one 2×2 subsquare, and it takes $O(n^3)$ time to find it.) A "naive" subsquare checker works as follows. Choose any pair of cells in the same row and generate the smallest square containing them. This involves $O(n^2)$ amount of work for each pair. There are $O(n^3)$ pairs which means that the naive subsquare checker has $O(n^5)$.

(b) Is the question "Does this Latin rectangle have an N_2 completion?" NP-complete?

Acknowledgments. The authors would like to thank C. J. Colbourn for helpful discussions relating to this paper. The comments of the referees are also appreciated.

REFERENCFS

[1] L. D. ANDERSEN, *Completing partial Latin squares*, Mat. Fys. Medd. Danske. Vid. Selsk., 41(1985), pp. 23–69.

[2] L. D. ANDERSEN AND E. MENDELSOHN, *A direct construction for Latin Squares without proper subsquares*, Ann. Discrete Math., 15 (1982), pp. 27–53.

[3] R. H. F. DENNISTON, *Remarks on Latin Squares with no subsquares of order two*, Utilitas Math., 13 (1978), pp. 299–302.

[4] A. KOTZIG, C. C. LINDNER AND A. ROSA, *Latin squares with no subsquares of order two and disjoint Steiner triple systems*, Utilitas Math., 7 (1975), pp. 287–294.

[5] K. HEINRICH, *Latin squares with no proper subsquares*, J. Combin. Theory, Ser. A (1980), pp. 346–353.

[6] A. KOTZIG AND J. TURGEON, *On certain constructions for Latin squares with no subsquares of order two*, Discrete Math., 16 (1976), pp. 263–270.

[7] M. MCLEISH, *On the existence of Latin squares with no subsquares of order two*, Utilitas Math., 8 (1975), pp. 41–53.

[8] E. MENDELSOHN AND A. ROSA, *One-factorizations of the complete graph—A survey*, J. Graph Theory, 9 (1985), pp. 43–65.

[9] N. S. MENDELSOHN, Private communication, quoted in [5].

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

PROJECTIONALLY EXPOSED CONES*

GEORGE PHILLIP BARKER†, MICHAEL LAÏDACKER‡ AND GEORGE POOLE‡

Abstract. Programming problems, based on the objective function and types of constraints, may be classified as linear, nonlinear, discrete, integer, and Boolean, just to name a few. These programming problems represent special cases of the more general Abstract Convex Programming Problem given by: Find Min $\{f(x): g(x) \in -K, x \in \Omega\}$ where $\Omega \subseteq \mathbb{R}^n$ is convex, K is a convex cone, and f, g are convex functions. Characterizations of optimality to the Abstract Convex Programming Problem are of paramount importance in the investigation of optimization problems. A cone K in $\mathbb{R}^n$ is called projectionally exposed if for each face F of K there exists a projection P_F of $\mathbb{R}^n$ such that $P_F(K) = F$. In particular, it has been shown that when the constraint function g of the Abstract Convex Programming Problem takes values in a projectionally exposed cone, then certain multipliers, associated with optimality, may be chosen from a smaller set (see § 6 of [Borwein and Wolkowicz, J. Math. Anal. Appl., 83 (1981), pp. 495–530]). This suggests that a study of such cones is both applicable and intrinsically interesting. With this motivation, the authors have undertaken a project to characterize the cones of $\mathbb{R}^n$ which are projectionally exposed.

Key words. cones, exposed faces, projections

AMS(MOS) subject classifications. 15A48, 15A04, 90C25

Introduction. In connection with their study of the (abstract) convex program Borwein and Wolkowicz [2] introduce the notion of a projectionally exposed cone (see the definitions in the next section). In particular when the constraint function takes values in a projectionally exposed cone, then certain multipliers may be chosen from a smaller set (see [2, §6]). This suggests that a study of such cones may be both applicable and intrinsically interesting. Although Borwein and Wolkowicz do not restrict their cones to be either closed or pointed, we shall do so in order to simplify the initial study. It is hoped that later work can relax these assumptions.

1. Definitions. Let V be a finite dimensional real inner product space of dimension n. In the examples we shall take V to be $\mathbb{R}^n$ with the usual inner product. However, we shall use functional notation, fx, in place of (f, x). That is, we shall use the inner product to identify the dual space V^* of linear functionals with V. A (convex) cone K in V is a subset such that for any $x, y \in K$, $\alpha, \beta \geqq 0$ we have $\alpha x + \beta y \in K$. The cone K is *pointed* iff it contains no subspace (i.e., $K \cap (-K) = \{0\}$); it is *closed* iff K is closed in the natural topology of V; K is *full* iff it has nonempty interior.

If K is a cone, the subspace spanned by K is $K - K = \{x - y | x, y \in K\}$. Since K is full in its span we shall assume that K has nonempty interior. We shall also assume that K is closed. This is a significant restriction and it is hoped that in subsequent work this assumption can be relaxed. Finally, we shall work primarily with pointed cones, but this hypothesis will be made explicit in the statements of the results.

For a cone K the *positive dual* K^* is the set of all nonnegative linear functionals on K:

$$K^* = \{f | fx \geqq 0 \text{ all } x \in K\}.$$

When K is closed we have $K^{**} = K$. A *face* of K is a (convex) subcone F of K such that

$$x \in K, y \in F, \text{ and } y - x \in K \text{ imply } x \in F.$$

* Received by the editors April 8, 1985; accepted for publication (in revised form) March 19, 1986.

† Department of Mathematics, University of Missouri, Kansas City, Missouri 64110.

‡ Department of Mathematics, Lamar University, Beaumont, Texas 77710.

This is denoted by $F \lhd K$. If we introduce an order relation in V by $x \geqq 0$ iff $x \in K$, then a subcone F is a face of K iff $0 \leqq x \leqq y$ and $y \in F$ imply $x \in F$. When K is pointed the order is a partial order.

DEFINITION 1.1. A face F is *exposed* iff there is an $f \in K^*$ such that $F = \{x \in K | fx = 0\}$. The cone K is *facially exposed* iff every face of K is exposed.

DEFINITION 1.2. Let $F \lhd K$.

(a) F is p-*exposed* (*projectionally exposed*) iff there is a projection P such that $PK = F$. If every face is p-exposed we call K p-exposed.

(b) F is o.p.-*exposed* iff there is an orthogonal projection P such that $PK = F$. If every face of K is o.p.-exposed, then K is o.p.-exposed.

If $S \subseteq K$, then the intersection of all faces containing S is a face of K which we denote by $\varphi(S)$. When $S = \{x\}$ we write $\varphi(x)$ for simplicity. If $F \lhd K$ and $\langle F \rangle$ is the *linear span* of F, then dim F is defined to be dim $\langle F \rangle$. An extreme ray of K is a ray which is a face. If $0 \neq x \in F$ and F is a ray which is a face, we call x an *extremal* of K.

A special class of cones arises in studying the solvability of finite systems of inequalities. These are the polyhedral cones. A cone K is *polyhedral* iff it has a finite number of extreme rays. An equivalent condition (cf. [4]) is that K should have a finite number of maximal faces. A *maximal* (*proper*) *face* is, of course, a face different from K which is contained in no other face of K.

2. Results and examples.

Remark 2.1. Since we are assuming that V is an inner product space, then $K^* \subset V$. Consequently, the statement $K \subset K^*$ makes sense. When it holds we call K *subpolar.*

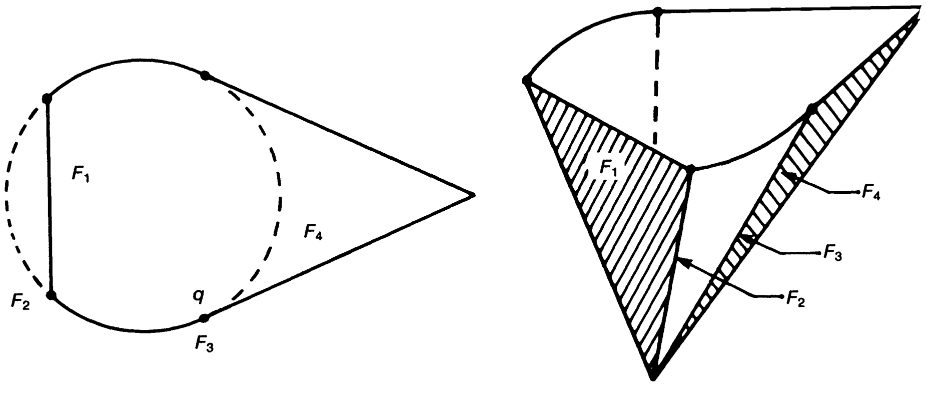
PROPOSITION 2.2. *Assume that closed full cone K is neither* $\{0\}$ *nor V. Then every extreme ray of K is* p-*exposed. If* $K \subset K^*$, *then every extreme ray is* o.p.-*exposed. Finally, if K is pointed and* o.p.-*exposed, then K is subpolar.*

Remark. We shall see in the examples following the proof that o.p.-exposed faces need not be exposed.

Proof. Let x be an extremal of K. Without loss we may take $V = \mathbb{R}^n$. Choose $f \in K^*$ such that $fx = 1$. If $K \subset K^*$ we can normalize x so that $x^Tx = 1$ where $x^Tx = (x, x)$. If we then take P to be the rank one projection $P = xf(P = xx^T$, respectively) where $xf(y) = (fy)x$ we have $PK = \varphi(x)$ so that P is the desired (orthogonal) projection. For the converse assume that K is pointed and that y is an extremal. If P is an orthogonal projection onto $\varphi(y)$, then $P = \alpha yy^T$ where $\alpha^{-1} = y^Ty > 0$. But $PK \subset K$ implies $y^T \in K^*$. Since K is the convex hull of its extreme rays, then $K \subseteq K^*$. □

Examples 2.3. Consider the cone in Fig. 1(b) whose cross section is given in Fig. 1(a). The line qr is tangent to the circle (as is the symmetric line) and $K = K^*$. The face $\varphi(q)$ is not exposed but is o.p.-exposed. Also the next theorem will show there are p-exposed cones which are not subpolar. The subpolar simplicial cone $K \subset \mathbb{R}^2$ of Fig. 2 shows that the conclusion of the proposition cannot be strengthened to $K = K^*$. Note that in this last example K^* is p-exposed but not o.p.-exposed.

For a closed pointed cone if both K and K^* are o.p.-exposed, then since $K = K^{**}$ we have $K = K^*$. But these are exactly the perfect cones [1] of which the nonnegative orthant and the positive semidefinite matrices are examples. The nonsimplicial perfect cones are thus (cf. [1]) nonpolyhedral examples of facially exposed cones (cf. our Definition 1.2(a) and [2, Def. 3.2]). Thus the only o.p.-exposed selfdual polyhedral cones are the images of the nonnegative orthant under an orthogonal matrix. However, there are perfect (i.e., o.p.-exposed selfdual) cones other than these examples which are homogeneous in the sense of [5]. An extensive discussion of selfdual cones in finite and infinite dimensional settings is found in [3].



(a) (b)

FIG. 1

THEOREM 2.4. *Let K be a pointed full polyhedral cone. Then K is* p-*exposed.*

Proof. We show first that there is a cone preserving projection onto any maximal face. Let $F_1, \cdots, F_m$ be the maximal faces of K, and let $f_i \in K^*$, $i = 1, \cdots, m$ be linear functionals such that

$$F_i = \{x \in K | f_i x = 0, i = 1, \cdots, m\}.$$

We construct a projection onto F_1. To this end set

$$K_0 = \{x \in V | f_1 x < 0, f_i x \geqq 0, i = 2, \cdots, m\}.$$

$K_0 \neq \emptyset$ since if x_0 is in the relative interior of $H_1 = \text{span } F_1$ and if $N(x_0)$ is an open ball around x_0 which meets no proper face of K other than H_1, then $N(x_0)$ meets K_0. So choose $p \in K_0$, let $L = \text{span } \{p\} = \{\alpha p | \alpha \in \mathbb{R}\}$ and let Q be the projection onto $\langle F_1 \rangle$ along L. Let $k \in K$. Then for some $\alpha_0 \in \mathbb{R}$ and $h \in \langle F_1 \rangle$ we have $k = h + \alpha_0 p$ and $h = Q(k)$. Now we have

$$h = k - \alpha_0 p, \qquad f_i(h) = f_i(k) - \alpha_0 f_i(p).$$

For $i = 1$ this becomes

$$0 = f_1(h) = f_1(k) - \alpha_0 f_1(p).$$

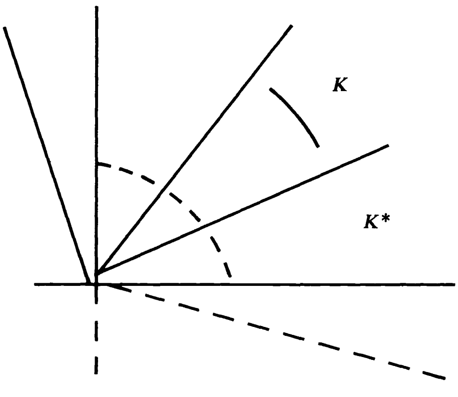


FIG. 2

But $f_1(k) \geqq 0, f_1(p) < 0$ so that $\alpha_0 \leqq 0$ or $-\alpha_0 \geqq 0$. Next for $i = 2, \cdots, m$ we see that $f_i(k) \geqq 0$ since $f_i \in K^*$ and $f_i(p) \geqq 0$ since $p \in K_0$. Thus $f_i(h) \geqq 0$, $i = 2, \cdots, m$. From the choice of the f_i this means that $h \in K$, hence $h \in F_1$. Since F_1 was an arbitrary maximal face we see that any maximal face is p-exposed.

Suppose $H \lhd K$. We choose a chain

$$H = H_0 \lhd H_1 \lhd \cdots \lhd H_p \lhd K,$$

where $\dim H_{j+1} = \dim H_j + 1$. Since each H_j satisfies the hypothesis and H_{j-1} is a maximal face of H_j there is a projection $\tilde{Q}_j$ of H_j onto H_{j-1} along some L_{j-1}. Each $\tilde{Q}_j$ can be extended to a projection Q_j of V onto $\langle H_{j-1}\rangle$ if we define Q_j to be zero on $L_{p-1} + \cdots + L_j$ and extend by linearity. Then $Q = Q_1$ is the desired projection. □

We continue the notation of the statement and proof of Theorem 2.4. We may normalize the f_i so that $|f_i| = 1$, where $|f_i|^2 = (f_i, f_i)$, for $i = 1, \cdots, m$. For each face F_i we define the complementary cone K_i by

$$K_i = \{x | f_i x < 0 \text{ and } f_j x \geqq 0 \text{ for } j \neq i\}.$$

With each pair of maximal support planes $\langle F_i\rangle$ and $\langle F_j\rangle$ we associate an angle $\theta \in (0, \pi)$ determined by

$$\cos\theta = -(f_i, f_j)$$

(see Fig. 3). In the proof of Theorem 2.4 we showed that each point of K_i determines a projection which takes K onto F_i. The next lemma shows that every cone preserving projection arises this way.

LEMMA 2.5. *Let P be a projection such that $P(K) = F_i$. Then* $\ker P \cap K_i \neq \{0\}$.

Proof. Since $\dim F_i = n - 1$ then $\dim \ker P = 1$. Let y be a nonzero vector in $\ker P$ which satisfies $f_i y < 0$. Then y and K_i lie in the same open halfspace determined by $\langle F_i\rangle$. We wish to show that $y \in K_i$. Assume not. Then for some $j \neq i, f_j y < 0$. Let $z \in F_j$ be nonzero so that $f_t z \geqq 0$ for $t \neq j$ and $f_j z = 0$. Further we may take $z \not\in F_i$. Let $Pz = w \in F_i$. Since $\ker P = \langle y\rangle$ we have $w = z + ky$. On the one hand,

$$0 = f_i w = f_i z + k f_i y$$

so that $k \geqq 0$. However, on the other hand we have

$$0 \leqq f_j w = f_j z + k f_j y = k f_j y$$

so that $k \leqq 0$. Thus $k = 0$ and $z = w \in F_i \cap F_j$, a contradiction. Thus $y \in K_i$. □

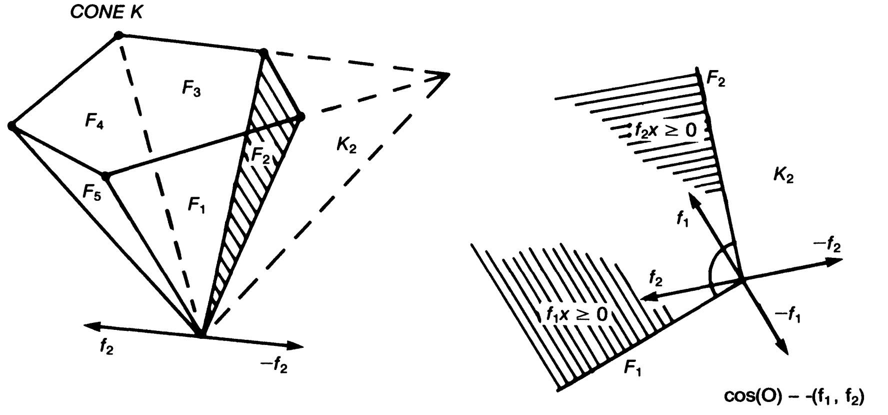


FIG. 3

THEOREM 2.6. *Let K be a pointed full polyhedral cone with maximal faces $F_1, \cdots, F_m$ and corresponding functionals $f_1, \cdots, f_m$. Then the face F_i is* o.p.*-exposed iff*

$$0 \leqq \theta = \cos^{-1}(-f_i, f_j) \leqq \pi/2 \quad \textit{for all } j \neq i.$$

Proof. Assume F_i is o.p.-exposed. Then by Lemma 2.5, $-f_i \in K_i$. Hence $-(f_i, f_j) \geqq 0$ for all $j \neq i$. Therefore $\cos^{-1}(-f_i, f_j) \leqq \pi/2$ for all $j \neq i$.

Conversely, if $\cos^{-1}(-f_i, f_j) \leqq \pi/2$ for all $j \neq i$, then $(-f_i, f_j) \geqq 0$. Also $(-f_i, f_i) < 0$, hence $-f_i \in K_i$. Thus the orthogonal projection onto $\langle F_1 \rangle$ takes K onto F_i. □

Let $\mathcal{V} = \{v_1, \cdots, v_n\}$ be a basis of V which consists of pairwise obtuse vectors, that is $(v_i, v_j) \leqq 0$ for all $i \neq j$. Apply the Gram–Schmidt process to obtain an orthonormal set $\mathcal{U} = \{u_1, \cdots, u_n\}$ defined by

$$u_1 = v_1/|v_1|, \quad w_1 = v_1, \qquad u_i = w_i/|w_i|, \quad |w_i|^2 = (w_i, w_i)$$

where

$$w_i = v_i - (v_i, u_1)u_1 - \cdots - (v_i, u_{i-1})u_{i-1} \quad \text{for } i = 2, \cdots, n.$$

THEOREM 2.7. *Let $\mathcal{V}$ and $\mathcal{U}$ denote the two sets of vectors in V previously described. Then for $k = 1, \cdots, n$*

(i) $v_k \in H(u_k) = \{x | (u_k, x) > 0\}$;

(ii) *there does not exist any vector w such that $(w, v_i) \leqq 0$ for $i = 1, \cdots, k$ and $w \in \bigcup_{i=1}^{k} H(u_i)$.*

Proof. We induct on k. The result is clear for $k = 1$ from the definition of $H(u_1)$. Suppose the result is true for all $k \leqq m < n$ and let $k = m + 1$. First note that by Bessel's inequality we have $(v_{m+1}, w_{m+1}) > 0$ so $v_{m+1} \in H(u_{m+1})$ which is (i). We shall establish part (ii) for $k = m + 1$ by contradiction. Thus suppose there is a w for which $(w, v_i) \leqq 0$ for $i = 1, \cdots, m + 1$ and $w \in \bigcup_{i=1}^{m+1} H(u_i)$. By the induction hypothesis with $k = m$ we must have $w \notin \bigcup_{i=1}^{m} H(u_i)$ so that $w \in H(u_{m+1})$. Thus $(w, u_{m+1}) > 0$, and in the Gram–Schmidt process we have

$$v_{m+1} = (v_{m+1}, u_1)u_1 + \cdots + (v_{m+1}, u_m)u_m + |w_{m+1}|u_{m+1}.$$

Taking the inner product with w yields

$$(*) \qquad (w, v_{m+1}) = (v_{m+1}, u_1)(w, u_1) + \cdots + (v_{m+1}, u_m)(w, u_m) + |w_{m+1}|(w, u_{m+1}).$$

The vectors in $\mathcal{V}$ are pairwise obtuse so in particular $(v_{m+1}, v_i) \leqq 0$ for $i = 1, 2, \cdots, m$ so that for these i we have $v_{m+1} \notin H(u_i)$ by the induction hypothesis. Consequently, for $i = 1, \cdots, m$ $(v_{m+1}, u_i)(w, u_i) \geqq 0$. But also $|w_{m+1}| > 0$ and $(w, u_{m+1}) > 0$, hence the sum on the right side of $(*)$ is strictly positive. This contradicts $(w, v_{m+1}) \leqq 0$, and the induction step is established. □

THEOREM 2.8. *Suppose $\mathcal{S} = \{w_1, w_2, v_1, v_2, \cdots, v_n\}$ is a subset of V for which the following are true:*

(i) *Any subset of $\mathcal{S}$ of cardinality n is linearly independent,*

(ii) *$V = \{v_1, \cdots, v_n\}$ and $U = \{u_1, \cdots, u_n\}$ satisfy the hypotheses of Theorem* 2.7.

(iii) $w_1, w_2 \notin \bigcup_{i=1}^{n} H(u_i)$.

Then $(w_1, w_2) > 0$.

Proof. First we show that neither w_1 nor w_2 lies in the hyperplane $L = \{x | (u_n, x) = 0\}$. Suppose $w_1 \in L$. Then from the orthogonality of the u_j we have that $w_1 \in L = \text{span}(u_1, \cdots, u_{n-1}) = \text{span}(v_1, \cdots, v_{n-1})$ which contradicts (i). Similarly, $w_2 \notin L$. Thus from (iii) we have for $i = 1, 2$, and $k = 1, \cdots, n - 1$ that $(w_i, u_k) \leqq 0$. Further, $(w_1, u_n) < 0$ and $(w_2, u_n) < 0$. Thus there are coefficients γ_i such that

$$w_2 = \gamma_1 u_1 + \gamma_2 u_2 + \cdots + \gamma_n u_n$$

where $\gamma_i \leqq 0$ ($i = 1, \cdots, n-1$) and $\gamma_n < 0$. Consequently,

$$(w_1, w_2) = \gamma_1(w_1, u_1) + \cdots + \gamma_n(w_1, u_n) > 0. \qquad \square$$

Let card $\mathcal{S}$ denote the cardinality of a (finite) subset of V. We may rephrase the conclusions of Theorems 2.7 and 2.8 as follows.

PROPOSITION 2.9. *Let $\mathcal{S}$ be a finite subset of V such that*

(i) *if $\mathcal{T} \subset \mathcal{S}$ and* card $\mathcal{T} \leqq n$, *then the vectors in $\mathcal{T}$ are linearly independent,*

(ii) *the vectors in $\mathcal{S}$ are pairwise obtuse.*

Then card $\mathcal{S} \leqq n + 1$.

THEOREM 2.10. *Let K be a pointed full polyhedral cone in V. If K is* o.p.*-exposed then K has $n =$* dim *V extreme rays.*

Remark. A closed pointed full polyhedral cone K with $n = \dim V$ extreme rays is called *simplicial.*

Proof. It suffices to show that K^* has exactly n extreme rays since K is simplicial iff K^* is simplicial. Suppose K^* has $k > n$ extreme rays. If H is a maximal face of K^* we can find $n-1$ linearly independent extremals in H, say $f_1, \cdots, f_{n-1}$. Let f_n be another extremal of K^* which is not in H. Let $L_j = \langle f_1, \cdots, f_{j-1}, f_{j+1}, \cdots, f_n \rangle$ be the subspace spanned by all the f_k except f_j. Then $\cup_{j=1}^{n} L_j$ is not a subspace so in particular since K^* has nonempty interior we can find an extremal f_{n+1} of K^* which is not in $\cup L_j$. Let $\{M_k\}$ be the collection of all subspaces spanned by subsets of cardinality $n-1$ of $\{f_1, \cdots, f_{n-1}\}$. Again there are only finitely many such subspaces so there is a $y \in$ interior K such that $y \not\in \cup M_k$. Let $\mathcal{S} = \{f_1, \cdots, f_{n+1}, -y\}$. From the construction of $\mathcal{S}$ we see that any subset of cardinality $\leqq n$ is linearly independent. Also as in the proof of Theorem 2.6 since $\ker f_k \cap K$ is a maximal face of K we have that $(f_i, f_j) \leqq 0$ for $i \neq j$. Also since $y \in$ interior K we have $-(y, f_k) = (-y, f_k) < 0$ for all k. But card $\mathcal{S} = n + 2$ which contradicts Proposition 2.9. Thus K^* and hence K has only n extremals. $\square$

COROLLARY 2.11. *The closed pointed full polyhedral cone K is* o.p.*-exposed if K is a subpolar simplicial cone.*

REFERENCES

[1] G. P. BARKER, *Perfect cones*, Linear Algebra Appl., 22 (1978), pp. 211–221.

[2] J. BORWEIN AND H. WOLKOWICZ, *Regularizing the abstract convex program*, J. Math. Anal. Appl., 83 (1981), pp. 495–530.

[3] B. IOCHUM, *Cônes autopolaires et algèbres de Jordan*, Lecture Notes in Mathematics No. 1049, Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1984.

[4] B.-S. TAM, *A note on polyhedral cones*, J. Austral. Math. Soc., 22 (1976), pp. 456–461.

[5] E. B. VINBERG, *The theory of convex homogeneous cones*, Trans. Moscow Math. Soc., 12 (1963), pp. 340–403.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

THE SPECTRA OF MATRICES HAVING SUMS OF PRINCIPAL MINORS WITH ALTERNATING SIGN*

JÜRGEN GARLOFF† AND VOLKER HATTENBACH†

Abstract. We present an observation on the localization of the spectrum of a matrix having sums of principal minors with alternating sign.

Key words. principal minors, PN-matrix, P-matrix, eigenvalues

AMS(MOS) subject classifications. 15A18, 15A57

In the past fifteen years, considerable attention has been paid in the economic literature to the class of the so-called PN-matrices and semi-PN-matrices (see [3], [6], [8, Chap. 7], [9]). A matrix is a (semi-) PN-matrix if every principal minor of odd order is positive, and every principal minor of even order is negative (provided that the order of the minors is greater than 1). We call a real matrix a SPN-matrix if all the sums of its principal minors of odd order are nonnegative and all the sums of its principal minors of even order are nonpositive. The class of the SPN-matrices obviously contains the PN-matrices as well as the nonnegative semi-PN matrices.

Example. Let the $n \times n$ matrix $A = (a_{ij})$ be defined by

$$a_{ij} = \begin{cases} 1 & \text{if } j \geqq i \\ a_j & \text{if } j < i \end{cases} \qquad (i,j = 1, \cdots, n),$$

see [7]. Then the principal minors of order $k+1$ are of the form $(1-a_{i_1})(1-a_{i_2}) \cdot \cdots \cdot (1-a_{i_k})$, where $1 \leqq i_1 < i_2 < \cdots < i_k \leqq n-1$. Thus, A is a PN-matrix and a SPN-matrix if and only if for all k, $a_k > 1$ and $a_k \geqq 1$, respectively.

The purpose of this note is to present an observation on the spectra of SPN-matrices. We note that a matrix A with the sign of the sums of its principal minors of order k equal to $(-1)^k$ or 0 can be transformed to a matrix having nonnegative sums of its principal minors by considering $-A$. Theorems concerning the spectra of such matrices may be found in [1].

Let $n \geqq 2$ and A be an $n \times n$ SPN-matrix. The characteristic polynomial of A is given by

$$p(x) = (-x)^n + s_1(-x)^{n-1} + s_2(-x)^{n-2} + \cdots - s_{n-1}x + s_n, \tag{1}$$

where s_k denotes the sum of the principal minors of order k of A. By definition, we have sign $s_k = (-1)^{k+1}$, $k = 1, \cdots, n$. Without loss of generality we may assume that A is nonsingular since otherwise we can divide $p(x)$ by x^μ, where μ is the multiplicity of the eigenvalue 0, to obtain a polynomial of lower degree whose coefficients have the same sign as the corresponding coefficients of $p(x)$. By using the companion matrix of p and the Perron–Frobenius theorem one obtains that A has a simple positive eigenvalue λ_1, say, equal to its spectral radius (see [4]).

Let the eigenvalues of A which are different from λ_1 be denoted by $\lambda_2, \cdots, \lambda_n$. It is easy to see that λ_2 is negative if $n = 2$. We therefore assume without loss of generality that $n \geqq 3$.

A matrix is called a P-matrix if all its principal minors are positive.

* Received by the editors August 7, 1985; accepted for publication (in revised form) April 21, 1986.

† Institute for Applied Mathematics, University of Freiburg i. Br., Freiburg i. Br., West Germany.

THEOREM. *Let A be a nonsingular* SPN-*matrix with spectral radius* λ_1. *Then* $-\lambda_2, \cdots, -\lambda_n$ *are the eigenvalues of a* P-*matrix.*

Proof. We divide the characteristic polynomial $p(x)$, cf. (1), of A by $x - \lambda_1$ and denote the resulting polynomial by $p_1(x)$. By the Horner scheme we obtain the following recurrence formula for the coefficients a_i of $p_1(x) = a_0x^{n-1} + a_1x^{n-2} + \cdots + a_{n-1}$

$$\begin{aligned} &a_0 = (-1)^n, \\ &a_k = a_{k-1}\lambda_1 + (-1)^{n-k}s_k, \qquad k = 1, \cdots, n-1. \end{aligned} \tag{2}$$

From the equality $a_{n-1}\lambda_1 + s_n = 0$ we conclude that sign $a_{n-1} = (-1)^n$ and by (2) recursively, sign $a_{n-k} = (-1)^n$, $k = 2, \cdots, n-1$. Hence by Vieta's formula, the kth elementary symmetric function σ_k of the eigenvalues $\lambda_2, \cdots, \lambda_n$,

$$\sigma_k(\lambda_2, \cdots, \lambda_n) = \sum_{2 \leqq i_1 < i_2 < \cdots < i_k \leqq n} \lambda_{i_1} \cdot \cdots \cdot \lambda_{i_k}, \qquad k = 1, \cdots, n-1,$$

has the sign $(-1)^k$. Then $\sigma_k(-\lambda_2, \cdots, -\lambda_n)$ is positive for $k = 1, \cdots, n-1$. By [1, Prop. 4] there exists a P-matrix such that $-\lambda_2, -\lambda_3, \cdots, -\lambda_n$ are the eigenvalues of this matrix. □

This theorem enables the use of the results on the localization of the spectra of P-matrices [1], [2], [5] in order to localize the spectra of SPN-matrices. The most important conclusions are given in the following corollary.

COROLLARY. *Let A be a nonsingular* SPN-*matrix with spectral radius* λ_1. *Then*

(i) $$|\arg \lambda_k| > \frac{\pi}{n-1}, \qquad k = 2, \cdots, n.$$

(ii) *There is at least one eigenvalue with negative real part; if there is exactly one such eigenvalue then*

$$|\arg \lambda_k| > \frac{\pi}{3}, \qquad k = 2, \cdots, n;$$

this bound is independent of n and cannot be improved.

(iii) *If $n > 2m + 3$ and there are exactly $m + 1$ eigenvalues with positive real parts or exactly m eigenvalues with negative real parts then there exists α satisfying*

$$|\arg \lambda_k| > \alpha > \frac{\pi}{n-1}, \qquad k = 2, \cdots, n.$$

REFERENCES

[1] D. HERSHKOWITZ, *On the spectra of matrices having nonnegative sums of principal minors*, Linear Algebra Appl., 55 (1983), pp. 81–86.

[2] D. HERSHKOWITZ AND A. BERMAN, *Localization of spectra of P- and P_0-matrices*, Linear Algebra Appl., 52/53 (1983), pp. 383–397.

[3] K.-I. INADA, *The production coefficient matrix and the Stolper–Samuelson condition*, Econometrica, 39 (1971), pp. 219–240.

[4] J. J. JOHNSON, *On partially non-positive matrices*, Linear Algebra Appl., 8 (1974), pp. 185–187.

[5] R. B. KELLOGG, *On complex eigenvalues of* M *and* P *matrices*, Numer. Math., 19 (1972), pp. 170–175.

[6] J. S. MAYBEE, *Some aspects of the theory of* PN-*matrices*, SIAM J. Appl. Math., 31 (1976), pp. 397–410.

[7] H. W. MILNES, *A note concerning the properties of a certain class of test matrices*, Math. Comp., 22 (1968), pp. 827–832.

[8] H. NIKAIDO, *Convex Structures and Economic Theory*, Academic Press, New York, 1968.

[9] Y. UEKAWA, M. C. KEMP AND L. L. WEGGE, P- *and* PN-*matrices, Minkowski- and Metzler-matrices, and generalizations of the Stolper–Samuelson and Samuelson–Rybczynski theorems*, J. Internat. Econom., 3 (1972), pp. 53–76.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 1, January 1987

010

MATRICES WITH SIGN SYMMETRIC DIAGONAL SHIFTS OR SCALAR SHIFTS*

DANIEL HERSHKOWITZ†, VOLKER MEHRMANN‡ AND HANS SCHNEIDER¶

Abstract. We generalize the concepts of sign symmetry and weak sign symmetry by defining k-sign symmetric matrices. For a positive integer k, we show that all diagonal shifts of an irreducible matrix are k-sign symmetric if and only if the matrix is diagonally similar to a Hermitian matrix. A similar result holds for scalar shifts, but requires an additional condition in the case $k = 1$. Extensions are given to reducible matrices.

Key words. matrix, Hermitian, sign symmetric, diagonal shift, scalar shift, diagonal similarity, graph, cordless circuit

AMS(MOS) subject classifications. 15A15, 15A57, 05C50

1. Introduction. A square complex matrix is said to be sign symmetric (weakly sign symmetric) if it has nonnegative products of symmetrically located minors (almost principal minors) (for detailed definition see Definition 2.11).

Weakly sign symmetric matrices were studied first by Gantmacher and Krein [8, p. 111] and by Koteljanskii [13]. That is why these matrices are also called GKK-matrices, e.g., Fan [5]. One reason for the interest in these classes of matrices is that they contain the important classes of the Hermitian matrices, the totally nonnegative matrices and the M-matrices. Another reason is the strong linkage between weak sign symmetry and the Fischer–Hadamard determinantal inequalities. This connection is studied in Gantmacher and Krein [8], Koteljanskii [12], Carlson [1], Green [9] and Hershkowitz and Berman [10].

A sufficient condition for positivity of the principal minors of a weakly sign symmetric matrix in terms of leading principal minors is given by Koteljanskii [13].

Relations between weakly sign symmetric matrices and ω-matrices are discussed in Engel and Schneider [4] and in Hershkowitz and Berman [11].

Sign symmetry and weak sign symmetry are also related to stability. It was proved by Carlson [2] that sign symmetric matrices whose principal minors are positive are stable, i.e., their spectra lie in the open right half plane. The same result is conjectured to hold for weakly sign symmetric matrices too.

In this paper we generalize the concepts of sign symmetry and weakly sign symmetry. We define k-sign symmetric matrices, where k is a nonnegative integer (see Definition 2.11). In view of our definition an $n \times n$ sign symmetric matrix is a k-sign symmetric matrix whenever $k \geqq (n - 1)/2$. The 1-sign symmetric matrices are those weakly sign symmetric matrices whose principal minors are real. Since reality of principal minors is assumed in all the results on weakly sign symmetric matrices quoted above, one may as well consider those as assertions on 1-sign symmetric matrices.

* Received by the editors July 22, 1985; accepted for publication April 21, 1986.

† Department of Mathematics, University of Wisconsin, Madison, Madison, Wisconsin 53706, and Department of Mathematics, Technion–Israel Institute of Technology, Haifa 32000, Israel. The work of this author was supported by the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel, grant 85-00153.

‡ Fakultät für Mathematik, Universität Bielefeld, Postfach 8640, D-4800 Bielefeld 1, Federal Republic of Germany. This research was carried out while this author was visiting the University of Wisconsin, Madison, Madison, Wisconsin 53706.

¶ Department of Mathematics, University of Wisconsin, Madison, Madison, Wisconsin 53706. The work of this author was supported by National Science Foundation grant DMS-8340189, Office of Naval Research grant N00014-85-K-1613, and by the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel, grant 85-00153.

After giving graph theoretic preliminaries in § 3, we characterize in § 4 the matrices all of whose diagonal shifts are k-sign symmetric, that is matrices A such that $A + D$ is k-sign symmetric for every real diagonal matrix D. Given a positive k, we show that an irreducible matrix satisfies this condition if and only if it is diagonally similar to a Hermitian matrix. Thus, a matrix satisfies the above shift condition for some positive k if and only if it satisfies the condition for every positive k.

For $k \geqq 2$, we prove in § 5 a similar result for a matrix A all of whose scalar shifts $A + tI$, where t is real, are k-sign symmetric. If $k = 1$ then we need an additional graph theoretic hypothesis, namely the reversibility of the chordless directed circuits of even length in the directed graph of A.

The extensions of our results to reducible matrices follow from a theorem in § 6 that a matrix A is k-sign symmetric if and only if every diagonal block in the Frobenius normal form of A is k-sign symmetric.

2. Definitions and notation.

Notation 2.1. We denote

$|\alpha|$: the cardinality of a set α.
$\mathbb{R}$: the field of real numbers.
$\mathbb{C}$: the field of complex numbers.
$[x]$: the maximal integer which is less than or equal to the real number x.

Notation 2.2. For a positive integer n we denote

$\langle n \rangle$: the set $\{1, 2, \cdots, n\}$.
$F^{n,n}$: the set of all $n \times n$ matrices over a field F.

Notation 2.3. For a (nondirected, simple) graph Γ we denote

$V(\Gamma)$: the vertex set of Γ.
$E(\Gamma)$: the edge set of Γ.
$[i, j]$: an edge between i and j, $i, j \in V(\Gamma)$. Observe that $[i, j] = [j, i]$.

DEFINITION 2.4. Let Γ be a graph. A sequence of edges in Γ which leads from i to j, $[i, p_1]$, $[p_1, p_2]$, $\cdots$, $[p_{m-1}, p_m]$, $[p_m, j]$, is called a *path* in Γ between i and j and is denoted by $[i, p_1, p_2, \cdots, p_m, j]$. A path $[i_1, \cdots, i_l]$ in Γ is said to be a *closed path* if $i_l = i_1$. A closed path $[i_1, \cdots, i_k, i_1]$ is said to be a *circuit* if $i_1, \cdots, i_k$ are distinct. A circuit is said to be of length k, or a *k-circuit*, if it consists of k edges.

Notation 2.5. For a (simple) directed graph (or digraph) Δ we denote

$V(\Delta)$: the vertex set of Δ.
$E(\Delta)$: the arc set of Δ.
(i, j): an arc from i to j, $i, j, \in V(\Delta)$. Observe that $(i, j) = (j, i)$ if and only if $i = j$.

DEFINITION 2.6. Let Δ be a digraph. A sequence of arcs in Δ from i to j, (i, p_1), (p_1, p_2), $\cdots$, (p_{m-1}, p_m), $(p_{m,j})$, is called a *directed path* in Δ from i to j and is denoted by $(i, p_1, p_2, \cdots, p_m, j)$. A directed path $(i_1, \cdots, i_l)$ in Δ is said to be a *closed directed path* if $i_l = i_1$. A closed directed path $(i_1, \cdots, i_k, i_1)$ is said to be a *directed circuit* (or *dicircuit*) if $i_1, \cdots, i_k$ are distinct. A dicircuit is said to be of length k, or a *k-dicircuit*, if it consists of k arcs.

DEFINITION 2.7. A digraph Δ is said to be *strongly connected* if either $|V(\Delta)| = 1$ or for every $i, j \in V(\Delta)$ there exists a directed path in Δ from i to j.

DEFINITION 2.8. A dicircuit $(i_1, \cdots, i_k, i_1)$, $k \geqq 3$, in a digraph Δ is said to have a *chord* if $E(\Delta)$ contains an arc (i_l, i_t) where

$$t \notin \begin{cases} \{l-1, l+1\}, & 1<l<k, \\ \{2,k\}, & l=1, \\ \{k-1,1\}, & l=k. \end{cases}$$

A dicircuit of length greater than 2 in Δ is said to be *chordless* if it has no chord.

DEFINITION 2.9. (i) A directed path $\alpha = (i_1, \cdots, i_k)$ in a digraph Δ is said to be *reversible* in Δ if $(i_k, \cdots, i_1)$ is also a directed path in Δ. In this case we denote the directed path $(i_k, \cdots, i_1)$ by α^*.

(ii) A digraph Δ is said to be *reversible* or *symmetric* if every directed path in Δ is reversible. Observe that Δ is reversible if and only if

$$(i,j) \in E(\Delta) \Rightarrow (j,i) \in E(\Delta).$$

Notation 2.10. Let A be an $n \times n$ matrix and let $\alpha, \beta \subseteq \langle n \rangle$, $\alpha, \beta \neq \phi$. We denote

$A[\alpha|\beta]$: the submatrix of A whose rows are indexed by α and whose columns are indexed by β in their natural orders.

$A[\alpha] = A[\alpha|\alpha]$,

$A(\alpha|\beta) = A[\langle n \rangle \backslash \alpha | \langle n \rangle \backslash \beta]$,

$A(\alpha) = A(\alpha|\alpha)$.

DEFINITION 2.11. (i) Let $A \in \mathbb{C}^{n,n}$ and let $\alpha, \beta \subseteq \langle n \rangle$, $|\alpha| = |\beta| > 0$. The submatrix $A[\alpha|\beta]$ of A is said to have *dispersion k* whenever $k = |\alpha| - |\alpha \cap \beta|$ (see also [12]). Submatrices with dispersion 1 are called *almost principal* submatrices.

(ii) Let k be a nonnegative integer. A square matrix A is said to be *k-sign symmetric* if it satisfies

$$\det A[\alpha|\beta] \det A[\beta|\alpha] \geqq 0 \tag{2.12}$$

for all submatrices $A[\alpha|\beta]$ of A with dispersion less than or equal to k. The set of all k-sign symmetric matrices in $\mathbb{C}^{n,n}$ is denoted by $SS^k_{\langle n \rangle}$.

(iii) A square matrix is called *sign symmetric* if (2.12) holds for all square submatrices $A[\alpha|\beta]$ of A (see also [13]). The set of all sign symmetric matrices in $\mathbb{C}^{n,n}$ is denoted by $SS_{\langle n \rangle}$.

(iv) A square matrix is called *weakly sign symmetric* if (2.12) holds for all submatrices $A[\alpha|\beta]$ of A with dispersion exactly 1 (see also [13]). The set of all weakly sign symmetric matrices in $\mathbb{C}^{n,n}$ is denoted by $WSS_{\langle n \rangle}$.

Remark 2.13. (i) Observe that for nonnegative integers k and m, the inequality $m > k$ implies $SS^m_{\langle n \rangle} \subseteq SS^k_{\langle n \rangle}$.

(ii) Let $\alpha, \beta \subseteq \langle n \rangle$, $|\alpha| = |\beta| > 0$, and let $k = |\alpha| - |\alpha \cap \beta|$. Since

$$|\alpha| + |\beta| - |\alpha \cap \beta| = |\alpha \cup \beta| \leqq n$$

and since $k \leqq |\alpha|$ it follows that $k \leqq n/2$. Thus, the dispersion of a square submatrix of an $n \times n$ matrix cannot exceed $n/2$. It now follows that for a nonnegative integer m, $m \geqq (n-1)/2$ we have $SS^m_{\langle n \rangle} = SS_{\langle n \rangle}$.

(iii) Since submatrices of a given matrix have dispersion 0 if and only if they are principal submatrices, it follows from Definition 2.11(ii) that the 0-sign symmetric matrices are just the matrices all of whose principal minors are real. Also, a k-sign symmetric matrix has real principal minors for every positive integer k.

(iv) Observe that $SS^1_{\langle n \rangle}$ is the set of those matrices in $WSS_{\langle n \rangle}$ that have real principal minors.

DEFINITION 2.14. Let A be an $n \times n$ matrix. The *graph* $\Gamma(A)$ of A and the *digraph* $\Delta(A)$ of A are defined by

$$V(\Gamma(A)) = V(\Delta(A)) = \langle n \rangle,$$

$$E(\Gamma(A)) = \{[i,j], i,j \in \langle n \rangle : a_{ij} \neq 0 \text{ or } a_{ji} \neq 0\},$$

$$E(\Delta(A)) = \{(i,j), i,j \in \langle n \rangle : a_{ij} \neq 0\}.$$

DEFINITION 2.15. Let A be an $n \times n$ matrix and let $\alpha = (i_1, \cdots, i_k)$ be a directed path in $\Delta(A)$. The corresponding *path product* is defined to be

$$\Pi_\alpha(A) = \prod_{j=1}^{k-1} a_{i_j i_{j+1}}.$$

DEFINITION 2.16. An $n \times n$ matrix A is said to be *combinatorially symmetric* if $\Delta(A)$ is reversible.

DEFINITION 2.17. Let $A.B. \in \mathbb{C}^{n,n}$. The matrices A and B are said to be *diagonally similar* if there exists a nonsingular diagonal matrix D such that

$$B = D^{-1}AD.$$

The matrices A and B are said to be *permutationally similar* if there exists a permutation matrix P such that

$$B = P^T AP.$$

DEFINITION 2.18. (i) A square matrix A is said to be in *Frobenius normal form* if A may be written in the block form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ 0 & A_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & A_{kk} \end{bmatrix}$$

where A_{ii} is an irreducible square matrix, $i = 1, \cdots, k$.

(ii) Let $A, B \in \mathbb{C}^{n,n}$. The matrix B is said to be a *Frobenius normal form of A* if B is in Frobenius normal form and if A and B are permutationally similar.

Remark 2.19. Observe that by Definition 2.18 the Frobenius normal form of a square matrix A is unique up to permutation similarity, and so Frobenius normal forms of A have the same diagonal blocks up to permutation similarity. Also, since, as is well known, a square matrix is irreducible if and only if its digraph is strongly connected, it follows that the diagonal blocks of the Frobenius normal form of A are the principal submatrices of A that correspond to the maximal strongly connected subgraphs (components) of $\Delta(A)$.

DEFINITION 2.20. Let $A \in \mathbb{C}^{n,n}$. A *diagonal shift* of A is a matrix $A + D$ where D is a real diagonal $n \times n$ matrix. A *scalar shift* of A is a matrix $A + tI$ where t is a real number.

3. Reversible digraphs.

PROPOSITION 3.1. *Let Δ be a digraph. Then every dicircuit in Δ is reversible if and only if every chordless dicircuit in Δ is reversible.*

Proof. The "only if" part is trivial. Conversely we prove our assertion by induction on the length of the dicircuits. Clearly, all dicircuits in Δ of length 1 and 2 are reversible. Also all 3-dicircuits are chordless and hence reversible. Assume that all dicircuits in Δ of length less than n ($n > 3$) are reversible, and let $\alpha = (i_1, \cdots, i_n, i_1)$ be an n-dicircuit in Δ. If α is chordless then it is reversible by the conditions of the proposition. Assume that α is not chordless. Without loss of generality we may assume that $(i_1, i_l) \in E(\Delta)$ where $l \neq 1, 2, n$. Observe that $\beta = (i_1, i_l, i_{l+1}, \cdots, i_n, i_1)$ is a dicircuit in Δ of length less than n and therefore, by the inductive assumption, β is reversible. Thus we have

$$(i_k, i_{k-1}) \in E(\Delta), \qquad k = l+1, \cdots, n, \tag{3.2}$$

$$(i_1, i_n) \in E(\Delta), \tag{3.3}$$

and

$$(i_l, i_1) \in E(\Delta). \tag{3.4}$$

By (3.4), $\gamma = (i_1, \cdots, i_l, i_1)$ is also a dicircuit in Δ. Since the length of γ is less than n, it follows from the inductive assumption that γ is reversible. Hence we have

$$(i_k, i_{k-1}) \in E(\Delta), \qquad k = 2, \cdots, l. \tag{3.5}$$

It now follows from (3.2), (3.3) and (3.5) that the dicircuit α is reversible. □

COROLLARY 3.6. *Let Δ be a strongly connected digraph. Then Δ is reversible if and only if every chordless dicircuit in Δ is reversible.*

Proof. The "only if" part is again trivial. Conversely, since Δ is strongly connected it follows that every arc (i, j) of Δ lies on some dicircuit α in Δ. By Proposition 3.1 the dicircuit α is reversible and hence $(j, i) \in E(\Delta)$. □

COROLLARY 3.7. *Let $A \in \mathbb{C}^{n,n}$. Then every diagonal block in the Frobenius normal form of A is combinatorially symmetric if and only if every chordless dicircuit in $\Delta(A)$ is reversible.*

Proof. Our claim follows immediately from Corollary 3.6 and Remark 2.19. □

4. Irreducible matrices with sign symmetric diagonal shifts.

LEMMA 4.1. *Let $A \in \mathbb{C}^{n,n}$ be diagonally similar to a Hermitian matrix. Then $A \in SS^k_{\langle n \rangle}$ for every nonnegative integer k.*

Proof. Let D be a diagonal matrix and B be a Hermitian matrix such that

$$A = D^{-1}BD.$$

For all $\alpha, \beta \subseteq \langle n \rangle$, $|\alpha| = |\beta| > 0$ we have

$$\begin{aligned} &\det A[\alpha|\beta] \det A[\beta|\alpha] \\ &\quad = \det D[\alpha] \det B[\alpha|\beta] \det D^{-1}[\beta] \det D[\beta] \det B[\beta|\alpha] \det D^{-1}[\alpha] \\ &\quad = \det B[\alpha|\beta] \det B[\beta|\alpha] = \det B[\alpha|\beta] \overline{\det B[\alpha|\beta]} \geqq 0. \end{aligned}$$ □

LEMMA 4.2. *Let $a, b \in \mathbb{C}$ and let*

$$p(t) = (t+a)(t+b).$$

(i) *If $p(t_1)$, $p(t_2) \in \mathbb{R}$ for two distinct real numbers t_1 and t_2 then either $a = \bar{b}$ or $a, b \in \mathbb{R}$.*

(ii) *If $b > a$ then $p(t) < 0$ for all t, $-b < t < -a$.*

Proof. (i) If $p(t_1)$, $p(t_2) \in \mathbb{R}$ for two distinct real numbers t_1 and t_2 then necessarily $a + b$, $ab \in \mathbb{R}$. Therefore, $p(t)$ is a polynomial with real coefficients. Since the roots of $p(t)$ are $-a$ and $-b$ our claim follows.

(ii) Immediate, since for $-b < t < -a$ we have $t + a < 0$ and $t + b > 0$. □

COROLLARY 4.3. *Let $a, b \in \mathbb{C}$. If $(t + a)(t + b) \geqq 0$ for all $t \in \mathbb{R}$ then $a = \bar{b}$.*

Proof. By Lemma 4.2(i) we have either $a = \bar{b}$ or $a, b \in \mathbb{R}$. In the latter case, by Lemma 4.2(ii) we have $a = b$. Hence, in each case, $a = \bar{b}$. □

In the following results we discuss k-sign symmetric matrices, $k \geqq 1$. As observed in Remark 2.13(iii), a matrix A is 1-sign symmetric if and only if A is weakly sign symmetric matrix with real principal minors. Note that a matrix $A \in \mathbb{C}^{n,n}$ may have nonreal principal minors even if all its diagonal shifts are in $WSS_{\langle n\rangle}$. This assertion can easily be verified for $n = 1, 2$. However it holds for higher orders too as demonstrated by the following irreducible 3×3 matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & i & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The following theorem relates weakly sign symmetric matrices to 1-sign symmetric matrices.

THEOREM 4.4. *Let $A \in \mathbb{C}^{n,n}$ be a weakly sign symmetric matrix and suppose that all the principal submatrices of A of order less than or equal to $n - 2$ are nonsingular. Then A has real principal minors if and only if the diagonal entries of A are real.*

Proof. The "only if" direction is obvious. Conversely, assume that A is a weakly sign symmetric matrix with real diagonal entries and nonsingular principal submatrices of order less than or equal to $n - 2$. We prove that the principal minors A are real by induction on the order of A. The claim is clear for matrices in $WSS_{\langle 1\rangle}$ and $WSS_{\langle 2\rangle}$. Assume it holds for weakly sign symmetric matrices of order less than n, $n \geqq 3$, and let $A \in WSS_{\langle n\rangle}$. Since every principal submatrix of a weakly sign symmetric matrix is also weakly sign symmetric, it follows from the inductive assumption that all principal minors of A of order less than n are real. Thus, all we have to prove is that $\det A$ is real.

Let $\alpha_1 = \langle n\rangle\backslash\{n\}$ and $\alpha_2 = \langle n\rangle\backslash\{n - 1\}$, and define a 2×2 matrix B by

$$b_{ij} = \det A[\alpha_i|\alpha_j], \qquad i,j = 1,2.$$

Since $A \in WSS_{\langle n\rangle}$ it follows that $B \in WSS_{\langle 2\rangle}$. Furthermore, b_{11} and b_{22} are principal minors of A of order $n - 1$, and hence b_{11} and b_{22} are real by the inductive assumption. Therefore, the determinant of B is real. By Sylvester's identity, e.g., [7, Vol. I, p. 33], we have

$$\det B = \det A[\langle n-2\rangle] \det A. \tag{4.5}$$

Since $\det A[\langle n - 2\rangle] \neq 0$ and by the inductive assumption $\det A[\langle n - 2\rangle]$ is real, it now follows from (4.5) that $\det A$ is real. □

The assumption of nonsingularity of the principal minors of A cannot be dropped from Theorem 4.4 as demonstrated by the matrix

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & i \\ 2 & -i & 0 \end{bmatrix}.$$

It is easy to verify that $A \in WSS_{\langle 3\rangle}$. However, $\det A = i$.

LEMMA 4.6. *Let $A \in \mathbb{C}^{n,n}$, $n \geqq 3$. Assume that α is an n-dicircuit in $\Delta(A)$ and that $\Gamma(A)$ consists of a single circuit. If $A + D \in SS^1_{\langle n\rangle}$ for all real diagonal matrices D then*

$$\Pi_\alpha(A) = \overline{\Pi_{\alpha^*}(A)}.$$

Proof. Without loss of generality assume that $\alpha = (1, \cdots, n, 1)$. Notice that $\Gamma(A)$ consists of the single circuit $[1, \cdots, n, 1]$. Since $A + D \in SS^1_{\langle n\rangle}$ for all real diagonal matrices D, it follows that

$$\begin{aligned} y(D) &= [\det(A+D)(1|2)][\det(A+D)(2|1)] \\ &= [a_{21}z(D)+(-1)^{n-2}a_{n1}p][a_{12}z(D)+(-1)^{n-2}a_{1n}q] \geqq 0 \end{aligned} \tag{4.7}$$

where

$$z(D) = \det(A+D)(1,2),$$

$$p = \prod_{j=2}^{n-1} a_{j,j+1},$$

and

$$q = \prod_{j=2}^{n-1} a_{j+1,j}.$$

Since α is a dicircuit in $\Delta(A)$, it follows that

$$a_{12}, p, a_{n1} \neq 0.$$

Since, as observed in Remark 2.13(iii), A has real principal minors, it follows that $z(D)$ attains every real value for suitable choices of D. Therefore, if $a_{21} = 0$ then for an appropriate choice of D we have $y(D) < 0$, which contradicts (4.7). Thus we have $a_{21} \neq 0$. Similarly we show that $a_{j+1,j} \neq 0, j = 1, \cdots, n-1$ and $a_{1n} \neq 0$. Since $A \in SS^1_{\langle n \rangle}$ we now have $a_{12}a_{21} > 0$. Dividing (4.7) by $a_{12}a_{21}$, we obtain

$$[z(D)+(-1)^{n-2}a_{n1}p/a_{21}][z(D)+(-1)^{n-2}a_{1n}q/a_{12}] \geqq 0. \tag{4.8}$$

Since $z(D)$ attains every real value, it follows from (4.8) and Corollary 4.3 that

$$a_{n1}p/a_{21} = \overline{a_{1n}q/a_{12}}. \tag{4.9}$$

Notice that since $a_{12}a_{21} > 0$ we have $a_{12}a_{21} = \overline{a_{12}a_{21}}$. Hence, by multiplying the left and the right sides of (4.9) by $a_{12}a_{21}$ and $\overline{a_{12}a_{21}}$, respectively, we obtain

$$\Pi_\alpha(A) = \overline{\Pi_{\alpha^*}(A)}. \qquad \square$$

LEMMA 4.10. *Let $A \in \mathbb{C}^{n,n}$ have real diagonal entries and assume that*

$$a_{ij}a_{ji} \in \mathbb{R} \quad \text{for all } i,j \in \langle n \rangle, \quad i \neq j. \tag{4.11}$$

If the equality

$$\Pi_\alpha(A) = \overline{\Pi_{\alpha^*}(A)} \tag{4.12}$$

holds for all chordless dicircuits α in $\Delta(A)$ then it holds for all dicircuits α in $\Delta(A)$.

Proof. Since A has real diagonal entries, it follows that (4.12) holds for 1-dicircuits. Also it follows from (4.11) that (4.12) holds for 2-dicircuits. Assume by induction that (4.12) holds for dicircuits of length less than m, $m \geqq 3$, and let $\alpha = (i_1, \cdots, i_m, i_1)$ be an m-dicircuit in $\Delta(A)$. If α is chordless then by the lemma's conditions (4.12) holds. If α is not chordless, then necessarily $m > 3$, and without loss of generality we may assume that $(i_1, i_l) \in E(\Delta(A))$, where $l \neq 1, 2, m$. Since $\Delta = \Delta(A[i_1, \cdots, i_m])$ is strongly connected and since by the conditions of the lemma every chordless dicircuit in Δ is reversible, it follows from Corollary 3.6 that Δ is reversible. Hence, $(i_l, i_1) \in E(\Delta(A))$ and hence $\beta = (i_1, i_l, i_{l+1}, \cdots, i_m, i_1)$ and $\gamma = (i_1, \cdots, i_l, i_1)$ are dicircuits in $\Delta(A)$ with length less than m. By the inductive assumption we have

$$\Pi_\beta(A) = \overline{\Pi_{\beta^*}(A)}, \tag{4.13}$$

and

$$\Pi_\gamma(A) = \overline{\Pi_{\gamma^*}(A)}. \tag{4.14}$$

Observe that

$$\Pi_\alpha(A)=\Pi_\beta(A)\Pi_\gamma(A)/a_{i_1i_l}a_{i_li_1}, \tag{4.15}$$

and

$$\Pi_{\alpha^*}(A)=\Pi_{\beta^*}(A)\Pi_{\gamma^*}(A)/a_{i_1i_l}a_{i_li_1}. \tag{4.16}$$

Since we have (4.11), it now follows from (4.13), (4.14), (4.15) and (4.16) that

$$\Pi_\alpha(A)=\overline{\Pi_{\alpha^*}(A)}. \qquad \square$$

We remark that Lemma 4.10 may be generalized. One can similarly prove the same conclusion under the assumptions that (4.11) holds and that (4.12) holds for all the dicircuits in an integral basis for the flow space of $\Delta(A)$, see [14].

THEOREM 4.17. *Let $A \in \mathbb{C}^{n,n}$ be an irreducible matrix and let k be a positive integer. Then the following are equivalent.*

(i) *$A + D \in SS^k_{\langle n\rangle}$ for all real diagonal matrices D.*

(ii) *The matrix A is diagonally similar to a Hermitian matrix.*

Proof. (i) ⇒ (ii). In view of Remark 2.13(i) it is enough to show this implicatior for $k = 1$. Assume that $A + D \in SS^1_{\langle n\rangle}$ for all real diagonal matrices D. Observe that since A is irreducible, the digraph $\Delta(A)$ is strongly connected. Let $\alpha = (i_1, \cdots, i_m, i_1)$ be a chordless m-dicircuit in $\Delta(A)$. By Definition 2.8 we have $m \geqq 3$. Let $B = A[i_1, \cdots, i_m]$. Notice that $\Gamma(B)$ consists of a single circuit. By Lemma 4.6 we have

$$\Pi_\alpha(A)=\overline{\Pi_{\alpha^*}(A)}. \tag{4.18}$$

It now follows from (4.18) that the chordless dicircuit α is reversible. By Corollary 3.6 the strongly connected digraph $\Delta(A)$ is reversible. Thus, since A is in $SS^1_{\langle n\rangle}$ it follows that

$$a_{ij}\neq 0 \Rightarrow a_{ij}a_{ji}>0 \quad \text{for all } i,j\in\langle n\rangle. \tag{4.19}$$

Furthermore, by Lemma 4.10 we have

$$\Pi_\alpha(A)=\overline{\Pi_{\alpha^*}(A)}, \tag{4.20}$$

for every dicircuit in $\Delta(A)$. Therefore, by Corollary 4.20 of [3] it follows from (4.19) and (4.20) that A is diagonally similar to a Hermitian matrix.

(ii) ⇒ (i). Assume that A satisfies (ii). Since $A + D$ is diagonally similar to a Hermitian matrix for all real diagonal matrices D, it follows by Lemma 4.1 that $A + D$ is in $SS^k_{\langle n\rangle}$. □

5. Irreducible matrices with sign symmetric scalar shifts. In this section we discuss matrices A all of whose scalar shifts are k-sign symmetric. Although the condition here is weaker than $A + D \in SS^k_{\langle n\rangle}$ for all real diagonal matrices D, the results are similar to those of the previous section.

The following lemma is well known and may be found in [8, p. 79, Remark 6^0].

LEMMA 5.1. *Let $A \in \mathbb{C}^{n,n}$ be a tridiagonal matrix such that*

$$a_{ii}\in\mathbb{R}, \qquad i=1, \cdots, n,$$

and

$$a_{i,i+1}a_{i+1,i}>0, \qquad i=1, \cdots, n-1.$$

Then A has distinct real eigenvalues. Furthermore, if $\lambda_1 < \cdots < \lambda_n$ are the eigenvalues of A and $\mu_1 < \cdots < \mu_{n-1}$ are the eigenvalues of $A(n)$ or of $A(1)$, then

$$\lambda_1<\mu_1<\lambda_2<\cdots<\mu_{n-1}<\lambda_n.$$

An easy well-known consequence of Lemma 5.1 is:

LEMMA 5.2. *Let $A \in \mathbb{C}^{n,n}$ be a tridiagonal matrix such that*

$$a_{ii} \in \mathbb{R}, \qquad i=1, \cdots, n,$$

and

$$a_{i,i+1}a_{i+1,i} \geqq 0, \qquad i=1, \cdots, n-1.$$

Then A has real eigenvalues.

LEMMA 5.3. *Let $A \in \mathbb{C}^{n,n}$, $n \geqq 3$, and suppose that if n is even then A is combinatorially symmetric. Assume that α is an n-dicircuit in $\Delta(A)$ and that $\Gamma(A)$ consists of a single circuit. If $A + tI \in SS^1_{\langle n \rangle}$ for all $t \in \mathbb{R}$ then*

$$\Pi_\alpha(A) = \overline{\Pi_{\alpha^*}(A)}.$$

Proof. Without loss of generality assume that $\alpha = (1, \cdots, n, 1)$. Notice that $\Gamma(A)$ consists of the single circuit $[1, \cdots, n, 1]$. Since $A + tI \in SS^1_{\langle n \rangle}$ for all $t \in \mathbb{R}$, it follows that

$$\begin{aligned} (5.4) \qquad f(t) &= \det(A+tI)(1|2)\det(A+tI)(2|1) \\ &= [a_{21}g(t)+(-1)^{n-2}a_{n1}p][a_{12}g(t)+(-1)^{n-2}a_{1n}q] \geqq 0 \quad \text{for all } t \in \mathbb{R}, \end{aligned}$$

where

$$g(t) = \det(A+tI)(1,2),$$

$$p = \prod_{j=2}^{n-1} a_{j,j+1},$$

and

$$q = \prod_{j=2}^{n-1} a_{j+1,j}.$$

Observe that since A has real principal minors, if n is odd then $g(t)$ attains every real value and our proof follows as the proof of Lemma 4.6 where (5.4), $f(t)$ and $g(t)$ replace (4.7), $y(D)$ and $z(D)$, respectively. If n is even then, since $a_{12} \neq 0$ and since A is combinatorially symmetric 1-sign symmetric matrix, it follows that $a_{12}a_{21} > 0$. Dividing (5.4) by $a_{12}a_{21}$, we obtain

$$(5.5) \qquad [g(t)+a][g(t)+b] \geqq 0 \quad \text{for all } t \in \mathbb{R}$$

where

$$a = \frac{a_{n1}p}{a_{21}}$$

and

$$b = \frac{a_{1n}q}{a_{12}}.$$

Since $g(t)$ attains infinitely many real values, it follows from Lemma 4.2(i) that either

$$(5.6) \qquad a = \bar{b},$$

or

$$(5.7) \qquad a, b \in \mathbb{R}.$$

If (5.6) holds, then we have (4.9) and we complete our proof as we do for Lemma 4.6.

If (5.6) does not hold, then we have (5.7) where $a \neq b$. Without loss of generality we may assume that

$$(5.8) \qquad b > a.$$

Observe tnat if $g(t)$ attains the value x then $g(t)$ attains every value which is greater than x. Thus, it follows from (5.5), (5.8) and Lemma 4.2(ii) that

$$(5.9) \qquad g(t) \geqq -a > -b \quad \text{for all } t \in \mathbb{R}.$$

Given that $A + tI \in SS^1_{\langle n\rangle}$ for all $t \in \mathbb{R}$ we have

$$(5.10) \qquad \begin{aligned} h(t) &= [\det (A+tI)(1|n)][\det (A+tI)(n|1)] \\ &= [a_{n1}r(t)+a_{21}q][a_{1n}r(t)+a_{12}p] \geqq 0 \quad \text{for all } t \in \mathbb{R} \end{aligned}$$

where

$$r(t) = \det (A+tI)(1,n).$$

Dividing (5.10) by the positive number $a_{1n}a_{n1}$, we obtain

$$(5.11) \qquad [r(t)+c][r(t)+d] \geqq 0 \quad \text{for all } t \in \mathbb{R}$$

where

$$c = \frac{a_{21}q}{a_{n1}},$$

and

$$d = \frac{a_{12}p}{a_{1n}}.$$

Observe that (5.8) implies that

$$(5.12) \qquad c > d.$$

As before, by Lemma 4.2(ii) it follows from (5.11) and (5.12) that

$$(5.13) \qquad r(t) \geqq -d > -c \quad \text{for all } t \in \mathbb{R}.$$

Observe that $(A + tI)(1, 2)$ and $(A + tI)(1, n)$ are tridiagonal matrices which satisfy the conditions of Lemma 5.1. Hence by Lemma 5.1 their eigenvalues are simple. Thus, for appropriate choices of t, the determinants of these matrices, which are $g(t)$ and $r(t)$ respectively, attain negative values. Hence, it follows from (5.9) and (5.13) that

$$(5.14) \qquad a, b, c, d > 0.$$

Let $\alpha_1 = \langle n\rangle \backslash \{1, n\}$ and $\alpha_2 = \langle n\rangle \backslash \{1, 2\}$, and define a 2×2 matrix B by

$$b_{ij} = \det (A+tI)[\alpha_i|\alpha_j], \qquad i,j = 1,2.$$

Observe that

$$(5.15) \qquad b_{11} = r(t), \quad b_{22} = g(t), \quad b_{12} = p, \quad b_{21} = q.$$

By Sylvester's identity we have

$$(5.16) \qquad \det B = [\det (A+tI)(1,2,n)][\det (A+tI)(1)] \quad \text{for all } t \in \mathbb{R}.$$

By Lemma 5.1 let λ be the minimal eigenvalue of $A(1, 2, n)$, and choose $t_0 = -\lambda$. Thus

$$(5.17) \qquad \det (A+t_0I)(1,2,n) = 0.$$

Furthermore, by Lemma 5.1 we have

$$r(t_0), g(t_0) < 0. \tag{5.18}$$

By (5.9), (5.13), (5.14) and (5.18) we now obtain

$$r(t_0)g(t_0) < ac = pq. \tag{5.19}$$

On the other hand, by (5.15), (5.16) and (5.17) we obtain

$$r(t_0)g(t_0) = pq,$$

which is a contradiction to (5.19). Therefore, our assumption that (5.6) does not hold is false, and our proof is completed. □

Lemma 5.3 does not hold for even n when we omit the combinatorial symmetry requirement as demonstrated by the following example.

Example 5.20. Let

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Let $\alpha, \beta \subseteq \langle 4 \rangle$, $|\alpha| = |\beta| = |\alpha \cap \beta| + 1$. To see that

$$\det(A + tI)[\alpha|\beta] \det(A + tI)[\beta|\alpha] \geqq 0 \quad \text{for all } t \in \mathbb{R}, \tag{5.21}$$

observe that the left side of (5.21) is equal to zero whenever $|\alpha| \leqq 2$, and is equal to t^2 whenever $|\alpha| = 3$.

We remark that it is possible that a condition which is somewhat weaker than combinatorial symmetry will do in Lemma 5.3.

However, for matrices with k-sign symmetric scalar shifts, $k > 1$, we do not need to state the condition of combinatorial symmetry.

LEMMA 5.22. *Let $A \in \mathbb{C}^{n,n}$, $n \geqq 3$, and let k be a positive integer, $k > 1$. Assume that α is an n-dicircuit in $\Delta(A)$ and that $\Gamma(A)$ consists of a single circuit. If $A + tI \in SS^k_{\langle n \rangle}$ for all $t \in \mathbb{R}$ then α is reversible in $\Delta(A)$.*

Proof. Without loss of generality assume that $\alpha = (1, \cdots, n, 1)$. Thus $\Gamma(A)$ consists of the single circuit $[1, \cdots, n, 1]$. Assume that α is not reversible. Without loss of generality we may assume that $a_{1n} = 0$. In view of Lemma 5.3 it is enough to consider the case where n is even. Hence we may assume that $n \geqq 4$. Recall that

$$A + tI \in SS^k_{\langle n \rangle} \quad \text{for all } t \in \mathbb{R} \tag{5.23}$$

yields that A has real principal minors. Also, it follows from (5.23), that

$$h(t) = \det(A + tI)(1|n) \det(A + tI)(n|1)$$
$$= [\tilde{p} + a_{n1} r(t)]\tilde{q} \geqq 0 \quad \text{for all } t \in \mathbb{R}$$

where

$$r(t) = \det(A + tI)(1, n),$$

$$\tilde{p} = \prod_{j=2}^{n} a_{j,j-1},$$

and

$$\tilde{q} = \prod_{j=2}^{n} a_{j-1,j}.$$

Observe that $h(t)$ is a polynomial in t of degree $n - 2$. Since it is nonnegative for all $t \in \mathbb{R}$ it follows that the leading coefficient $a_{n1}\tilde{q}$ must be nonnegative. In fact, since α is a dicircuit in $\Delta(A)$ we have

$$a_{n1}\tilde{q} > 0. \tag{5.24}$$

We distinguish between two cases:

Case 1. $n = 4$. By (5.23) we have

$$\begin{aligned} f(t) &= \det (A + tI)(1|2) \det (A + tI)(2|1) \\ &= [a_{21}g(t) + a_{41}a_{23}a_{34}]a_{12}g(t) \\ &= [a_{21}a_{12}g(t) + a_{41}\tilde{q}]g(t) \geqq 0 \quad \text{for all } t \in \mathbb{R} \end{aligned} \tag{5.25}$$

where

$$g(t) = \det (A + tI)(1,2).$$

If $a_{43} \neq 0$ then, since $a_{34} \neq 0$, $g(t)$ attains also negative values (for example for $t = -a_{33}$). Thus, in view of (5.24) we can choose t_0 such that $g(t_0) < 0$ and

$$|a_{21}a_{12}g(t_0)| < a_{41}\tilde{q}.$$

But then $f(t_0) < 0$ in contradiction to (5.25). Therefore we must assume that $a_{43} = 0$. Since $k > 1$ we now obtain by (5.23) that

$$\det (A + tI)(1,3|2,4) \det (A + tI)(2,4|1,3) = -a_{41}a_{23}a_{12}a_{34} \geqq 0,$$

which is a contradiction to (5.24).

Case 2. $n > 4$. By (5.23) we have

$$\begin{aligned} \tilde{f}(t) &= \det (A + tI)(1, n-1|2, n) \det (A + tI)(2, n|1, n-1) \\ &= [a_{21}a_{n,n-1}\tilde{g}(t) - a_{n1}\tilde{q}/a_{12}a_{n-1,r}][a_{12}a_{n-1,n}\tilde{g}(t)] \\ &= [a_{12}a_{21}a_{n-1,n}a_{n,n-1}\tilde{g}(t) - a_{n1}\tilde{q}]g(\tilde{t}) \geqq 0 \quad \text{for all } t \in \mathbb{R} \end{aligned} \tag{5.26}$$

where

$$\tilde{g}(t) = \det (A + tI)(1, 2, n-1, n).$$

By Lemma 5.2 $\tilde{g}(t)$ attains every nonnegative value. Thus, in view of (5.24) we can choose t_0 such that $\tilde{g}(t_0) > 0$ and

$$|a_{12}a_{21}a_{n-1,n}a_{n,n-1}\tilde{g}(t_0)| < a_{n1}\tilde{q}.$$

But then $\tilde{f}(t_0) < 0$ in contradiction to (5.26).

In each case we obtain a contradiction, which means that our assumption that α is not reversible is false. □

We now state the theorem for the irreducible case.

THEOREM 5.27. *Let $A \in \mathbb{C}^{n,n}$ be an irreducible matrix and let k be a positive integer, $k \geqq 2$. Then the following are equivalent.*

(i) *$A + tI \in SS^k_{\langle n \rangle}$ for all $t \in \mathbb{R}$.*

(ii) *$A + tI \in SS^1_{\langle n \rangle}$ for all $t \in \mathbb{R}$ and every chordless dicircuit in $\Delta(A)$ is reversible.*

(iii) *$A + tI \in SS^1_{\langle n \rangle}$ for all $t \in \mathbb{R}$ and every chordless dicircuit of even length in $\Delta(A)$ is reversible.*

(iv) *The matrix A is diagonally similar to a Hermitian matrix.*

Proof. (i) ⇒ (ii). Lemma 5.22 yields that every chordless dicircuit in $\Delta(A)$ is reversible. The rest of the implication is trivial.

(ii) ⇒ (iii). Obvious.

(iii) ⇒ (iv). The proof follows exactly as the proof of the part (i) ⇒ (ii) in Theorem 4.17, where D is replaced by tI, and where Lemma 5.3 is used instead of Lemma 4.6.

(iv) ⇒ (i). Since $A + tI$ is diagonally similar to a Hermitian matrix for all $t \in \mathbb{R}$, it follows by Lemma 4.1 that $A + tI \in SS^k_{\langle n\rangle}$. □

6. Reducible matrices with sign symmetric shifts.

THEOREM 6.1. *Let $A \in \mathbb{C}^{n,n}$ have the block form*

$$A=\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are square, and let k be a nonnegative integer. Then A is k-sign symmetric if and only if A_{11} and A_{22} are k-sign symmetric.

Proof. Clearly, if A is k-sign symmetric then so are A_{11} and A_{22}. Conversely, assume that A_{11} and A_{22} are k-sign symmetric and let $\alpha, \beta \subseteq \langle n\rangle$ be such that $q = |\alpha| = |\beta| > 0$ and

$$q - |\alpha \cap \beta| \leqq k. \tag{6.2}$$

We shall show that

$$\det A[\alpha|\beta] \det A[\beta|\alpha] \geqq 0. \tag{6.3}$$

Let m be the order of A_{11}. Denote by

$$\alpha' = \alpha \cap \langle m\rangle, \quad \alpha'' = \alpha \backslash \alpha', \quad \beta' = \beta \cap \langle m\rangle, \quad \beta'' = \beta \backslash \beta'.$$

Observe that

$$|a'| + |\alpha''| = |\beta'| + |\beta''| = q, \tag{6.4}$$

and hence

$$|\alpha'| + |\beta''| + |\beta'| + |\alpha''| = 2q. \tag{6.5}$$

In view of (6.5) we need to consider only the following two cases.

Case 1. $|\alpha'| + |\beta''| > q$ or $|\alpha''| + |\beta'| > q$. Assume that

$$|\alpha'| + |\beta''| > q. \tag{6.6}$$

By (6.4) we have $|\alpha'|, |\beta''| > 0$. Since $A[\beta''|\alpha'] = 0$ it follows from (6.6) by the easy direction of the Frobenius–König theorem [6] that $A[\beta|\alpha]$ is singular and hence

$$\det A[\alpha|\beta] \det A[\beta|\alpha] = 0.$$

Case 2. $|\alpha'| + |\beta''| = |\alpha''| + |\beta'| = q$. If $|\alpha'| = q$ $[|\beta''| = q]$ then $|\alpha''| = 0$ $[|\beta'| = 0]$ and hence $|\beta'| = q$ $[|\alpha''| = q]$. In this case $A[\alpha|\beta]$ and $A[\beta|\alpha]$ are submatrices of A_{11} $[A_{22}]$ and (6.3) follows. If $|\alpha'|, |\beta''| < q$ then observe that $A[\alpha|\beta]$ and $A[\beta|\alpha]$ are reducible. Furthermore, we have

$$\det A[\alpha|\beta] = \det A_{11}[\alpha'|\beta'] \det A_{22}[\alpha''|\beta''] \tag{6.7}$$

and

$$\det A[\beta|\alpha] = \det A_{11}[\beta'|\alpha'] \det A_{22}[\beta''|\alpha'']. \tag{6.8}$$

By (6.2), the sets α' and α'' contain at most k indices which are not in β' and β'', respectively. Hence, since A_{11} and A_{22} are k-sign symmetric, inequality (6.3) follows from (6.7) and (6.8). □

In view of Remark 2.13(ii) we obtain the following immediate corollary to Theorem 6.1.

COROLLARY 6.9. *Let $A \in \mathbb{C}^{n,n}$ have the block form*

$$A=\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are square. Then A is sign symmetric if and only if A_{11} and A_{22} are sign symmetric.

We remark that the "only if" part of Theorem 6.1 holds trivially also when we replace "k-sign symmetric" by "weakly sign symmetric." On the other hand, weak sign symmetry of A_{11} and A_{22} does not imply in general the weak sign symmetry of A for matrices with nonreal principal minors, as demonstrated by the following example.

Example 6.10. Let

$$A=\begin{bmatrix} i & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

where A_{11} is a 1×1 matrix. Obviously, the matrices A_{11} and A_{22} are weakly sign symmetric. However, the matrix A is not in $WSS_{\langle 3\rangle}$ since

$$\det A(3|2) \det A(2|3) = -1.$$

Since the class $SS^k_{\langle n\rangle}$ is invariant under permutation similarity, the following is a corollary to Theorem 6.1.

COROLLARY 6.11. *Let k be a nonnegative integer. A square matrix A is k-sign symmetric if and only if every diagonal block in the Frobenius normal form of A is k-sign symmetric.*

Let A be a square matrix. Observe that every dicircuit in $\Delta(A)$ is a dicircuit in $\Delta(B)$ where B is some diagonal block in the Frobenius normal form of A. Thus, the following theorem for the general case follows directly from Theorems 4.17 and 5.27 and Corollary 6.11.

THEOREM 6.12. *Let $A \in \mathbb{C}^{n,n}$ and let k and m be positive integers, $m \geqq 2$. Then the following are equivalent.*

(i) *$A + D \in SS^k_{\langle n\rangle}$ for all real diagonal matrices D.*

(ii) *$A + tI \in SS^m_{\langle n\rangle}$ for all $t \in \mathbb{R}$.*

(iii) *$A + tI \in SS^1_{\langle n\rangle}$ for all $t \in \mathbb{R}$ and every chordless dicircuit in $\Delta(A)$ is reversible.*

(iv) *$A + tI \in SS^1_{\langle n\rangle}$ for all $t \in \mathbb{R}$ and every chordless dicircuit of even length in $\Delta(A)$ is reversible.*

(v) *Every diagonal block in the Frobenius normal form of A is diagonally similar to a Hermitian matrix.*

REFERENCES

[1] D. CARLSON, *Weakly sign-symmetric matrices and some determinantal inequalities*, Colloq. Math., 27 (1967), pp. 123–129.

[2] ———, *A class of positive stable matrices*, J. Res. Nat. Bur. Standards, 78B (1974), pp. 1–2.

[3] G. M. ENGEL AND H. SCHNEIDER, *Cyclic and diagonal products on a matrix*, Linear Algebra Appl., 7 (1973), pp. 301–335.

[4] ———, *The Hadamard-Fischer inequality for a class of matrices defined by eigenvalue monotonicity*, Linear and Multilinear Algebra, 4 (1976), pp. 155–176.

[5] K. Fan, *Subadditive functions on a distributive lattice and an extension of Szász's inequality*, J. Math. Anal. Appl., 18 (1967), pp. 262–268.

[6] G. F. Frobenius, *Über zerlegbare Determinanten*, Sitzungsber. Preuss. Akad. Wiss., Berlin, 1917, pp. 274–277. See also: Ges. Abh., 3, Springer, Berlin–Heidelberg–New York, 1968, pp. 701–704.

[7] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[8] F. R. Gantmacher and M. G. Krein, *Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme*, Akademie Verlag, Berlin, 1960.

[9] B. M. Green, *On weakly sign-symmetric matrices*, Linear Algebra Appl., 8 (1974), pp. 355–360.

[10] D. Hershkowitz and A. Berman, *Necessary conditions and a sufficient condition for the Fischer-Hadamard inequalities*, Linear and Multilinear Algebra, 13 (1983), pp. 67–72.

[11] ———, *Notes on ω- and τ-matrices*, Linear Algebra Appl., 58 (1984), pp. 169–183.

[12] D. M. Koteljanskii, *The theory of nonnegative and oscillating matrices*, Amer. Math. Soc. Transl. Ser. 2, 27 (1963), pp. 1–8.

[13] ———, *A property of sign symmetric matrices*, Amer. Math. Soc. Transl. Ser. 2, 27 (1963), pp. 19–24.

[14] B. D. Saunders and H. Schneider, *Flows on graphs applied to diagonal similarity and diagonal equivalence for matrices*, Discrete Math., 24 (1978), pp. 205–220.

COMBINATORIAL CANONICAL FORM OF LAYERED MIXED MATRICES AND ITS APPLICATION TO BLOCK-TRIANGULARIZATION OF SYSTEMS OF LINEAR/NONLINEAR EQUATIONS*

KAZUO MUROTA†, MASAO IRI† AND MASATAKA NAKAMURA‡

Abstract. With a view to obtaining an efficient procedure for solving large-scale systems of equations, canonical block-triangular forms are defined for layered mixed matrices and for mixed matrices, and some practical examples are presented. The canonical forms are obtained from a straightforward application of the decomposition principle for submodular functions. The relation to the existing decomposition techniques for electrical networks, as well as to the Dulmage–Mendelsohn decomposition, is also discussed.

Key words. block-triangularization, layered mixed matrix, submodular function, Dulmage–Mendelsohn decomposition

AMS(MOS) subject classifications. 15A21, 05, 65, 68

1. Introduction. When solving a system of linear equations

$$A\mathbf{x}=\mathbf{b} \tag{1.1}$$

repeatedly for various values of the right-hand side vector $\mathbf{b} = \mathbf{b}(\theta)$ containing parameters θ, it is now standard to first decompose A (possibly with permutations of rows and columns) into LU-factors as

$$A=LU, \tag{1.2}$$

and then solve the triangular systems $L\mathbf{y} = \mathbf{b}$, $U\mathbf{x} = \mathbf{y}$ for different values of $\mathbf{b} = \mathbf{b}(\theta)$. It is most important here that the LU-factors of A can be determined independently of the parameters θ.

No less of interest are the cases where the coefficient A, as well as $\mathbf{b}$, changes with parameters, but with its zero/nonzero pattern kept fixed. Such situations often arise in practice, for example, in solving a system of nonlinear equations by the Newton method, or in determining the frequency characteristic of an electrical network by computing its responses to inputs of various frequencies. In this case we cannot calculate the LU-factors of A in advance, so that we usually resort to the so-called graph-theoretic methods and rearrange the equations and the variables to obtain a block-triangular form (see, e.g., [11], [21], [22], [24]). In particular, the block-triangularization based on the structure theory of bipartite graphs has proved to be effective, and is known as the Dulmage–Mendelsohn decomposition (abbreviated to *DM-decomposition*) [4], [5], [6], [7]. Then, each time the parameter values are specified, the equations corresponding to the DM-blocks may be solved either by direct inversion through LU-decomposition or by some iterative method.

The above two approaches, the LU-decomposition and the DM-decomposition, are two extremes in that the former admits arbitrary elementary row transformations on A and the latter restricts itself to permutations only. In other words, the LU-decomposition regards the entries of A as numbers belonging to a field in which arithmetic operations

* Received by the editors August 5, 1985; accepted for publication (in revised form) March 31, 1986.

† Department of Mathematical Engineering and Instrumentation Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

‡ Department of Natural and Artificial Systems, College of Arts and Sciences, University of Tokyo, Meguro-ku, Tokyo 153, Japan.

are defined, whereas the DM-decomposition treats them as if they were symbols, or indeterminates if one prefers algebraic terms. It is often the case, however, that part of the entries of A are to be regarded as numbers and the remaining as symbols.

To be more concrete, suppose a system of linear/nonlinear equations

$$\mathbf{f}(\mathbf{x})=\mathbf{0} \tag{1.3}$$

is to be solved by the Newton method. The equations may be divided into linear and nonlinear parts as

$$\mathbf{f}(\mathbf{x})=Q\mathbf{x}+\mathbf{g}(\mathbf{x}) \tag{1.4}$$

where Q is a constant matrix. Accordingly, the Jacobian matrix $J(\mathbf{x})$ of $\mathbf{f}(\mathbf{x})$ is expressed as

$$J(\mathbf{x})=Q+T(\mathbf{x}) \tag{1.5}$$

where $T(\mathbf{x})$ is the Jacobian matrix of $\mathbf{g}(\mathbf{x})$. Then we may regard the nonvanishing entries of $T(\mathbf{x})$ as independent symbols on which no arithmetic operations are expected, whereas the usual elimination operations could be defined for the matrix Q.

Another typical example is a system of equations describing an electrical network, which is made up of equations for conservation laws (i.e., Kirchhoff's laws) and those for element characteristics (see Example 3.1). The former, stemming from the topological incidence relations in the underlying graphs, involve only ±1 as the coefficients and hence are amenable to elimination operations. The latter, on the other hand, consist of coefficients which are contaminated by various noises and errors, and therefore may be modelled as independent transcendentals.

The present paper aims at establishing a decomposition technique for systems of linear/nonlinear equations such that the coefficients are classified into two groups as explained above. A canonical form is introduced for a matrix A of the form

$$A=\left(\begin{array}{c} Q \\ \hline T \end{array}\right) \tag{1.6}$$

where the entries of Q belong to a subfield $\mathbf{K}$ and the nonvanishing entries of T are transcendentals (in an extension field $\mathbf{F}$) which are algebraically independent over $\mathbf{K}$. A uniquely determined block-triangular form is obtained with the diagonal square blocks being nonsingular; for a singular A, rectangular blocks (corresponding to horizontal and vertical tails in the DM-decomposition) also appear, both being of full rank. The decomposition can be found by an efficient algorithm so that it can be applied to large-scale practical problems, of which some examples are given in §4.

The relations of the canonical form of this paper to the decompositions for electrical networks so far proposed (mentioned below), as well as to the combinatorial canonical form of a matrix with respect to its pivotal transforms introduced by Iri in [14], is also discussed in §5 and §7 with special reference to the admissible row transformations on matrices.

There have been several combinatorial studies on the rank of matrices in relation to splitting like (1.6); e.g., "2-block rank" of [13], matroidal characterization (see Theorem 3.1 below) of the rank of the matrix (1.6), and "Rank-Identity for mixed matrices" of [28] (see [18] for their relations).

In the literature on electrical network theory, it has been known that a system of equations describing an electrical network can be put into a block-triangular form if one chooses appropriate bases (tree-cotree pairs) for Kirchhoff's laws and rearranges the variables and the equations (for both Kirchhoff's laws and element characteristics). As far as the present authors know, the decomposition of a pair of current-graph and voltage-graph is investigated in [35], [36] in graph-theoretic terms for the networks involving controlled sources. Based on the result of [42], a decomposition of those networks which have admittance expressions was considered by Iri around 1979 [16] (see [17] for explicit illustration) using the notion of minimum-cover in an independent-matching problem. An incomplete attempt has been made by Nakamura and Iri [30], [31], [32] to define a block-triangularization for a system of equations describing the most general class of networks with arbitrary mutual couplings (such as those containing controlled sources, nullators and norators) following the theoretical framework ([15], [19], [20], [29], [33], [40], [41]) for the principal partition of matroids, or the decomposition of submodular functions.

Then it is shown by Murota in the unpublished report [25], which may be regarded as a preliminary version of the present paper, that (i) the method proposed by Nakamura and Iri in [30], [31], [32] produces too fine a partition for a useful block-triangularization of electrical networks, as will be demonstrated in §5 below; (ii) nevertheless, if one notices an appropriate identity characterizing the rank of the matrix (1.6), the basic idea of [30], [31], [32] can be modified to yield a block-triangularization for the electrical networks treated there; and moreover (iii) the modified method can be used in obtaining a block-triangularization for more general systems of equations like those mentioned above.

The canonical form defined in this paper has been obtained by establishing a new identity (Theorem 4.2) for the rank of a matrix of the form (1.6) and by applying the same decomposition principle as that in [30], [31], [32] to the relevant submodular function appearing in the identity. However, once the canonical form is found for a specific problem, it would be possible to describe it without explicit reference to submodularity or the decomposition principle for submodular functions. In fact, it could be described in a constructive manner in terms of an algorithm which is composed of pivoting operations on a matrix and path-searching in a graph. It should be emphasized, nevertheless, that the approach based on the general principle is heuristically effective, affords a proper perspective, clarifies the relation among various techniques for block-triangularization and suggests further meaningful extensions.

2. Preliminaries. Some results on the decomposition principle for submodular functions [15], [19], [20], [29], [33], [40], [41] are briefly summarized here for later references.

Let C be a finite set, and $p{:}2^C \rightarrow \mathbf{R}$ be a submodular function defined on it, i.e.,

$$p(X\cup Y)+p(X\cap Y)\leqq p(X)+p(Y) \tag{2.1}$$

for X, $Y \subset C$. (Throughout this paper, $X \subset C$ does not exclude $X = C$.) The family of those subsets of C which give the minimum of p will be denoted by $L(p)$:

$$L(p)=\{X|X\subset C, p(X)\leqq p(Y) \text{ for all } Y\subset C\}. \tag{2.2}$$

From the submodularity (2.1), it follows that

$$X\cup Y, X\cap Y\in L(p) \quad \text{for } X, Y\in L(p).$$

In other words, $L(p)$ is a (distributive) sublattice [2] of the Boolean lattice 2^C. Note that the length of a maximal chain in $L(p)$ from min $L(p)$ to max $L(p)$ is uniquely determined.

By the structure theory of distributive lattices [1], [2], there exists a one-to-one correspondence between sublattices of 2^C and partitions of C into partially ordered blocks. Furthermore, when a sublattice is derived from a submodular function as (2.2), "minors" are induced on the blocks. To be more specific, the following is known as (a version of)

Jordan–Hölder type theorem for submodular functions. (The proof is straightforward; see, e.g., [19].)

THEOREM 2.1. *Let p be a submodular function defined on a finite set C, and $L(p)$ the family of minimizers of p. Put $X_0 = \min L(p)$ and $X_r = \max L(p)$.*

(1) *Any maximal chain in $L(p)$*

$$X_0 \subsetneq X_1 \subsetneq \cdots \subsetneq X_r$$

determines a family of intervals (difference sets)

$$\{C_i | C_i = X_i \setminus X_{i-1}, i = 1, \cdots, r\},$$

which is independent of the choice of a maximal chain, and hence provides a unique partition of C into disjoint subsets (blocks)

$$\mathscr{P} = \{C_0; C_1, \cdots, C_r; C_\infty\}$$

where $C_0 = X_0$ and $C_\infty = C \setminus X_r$. ($C_0$ and/or C_∞ can be empty.)

(2) *The "minors" of p defined by*

$$p_i(Y) = p(X_{i-1} \cup Y) - p(X_{i-1}) \quad \textit{for } Y \subset C_i \tag{2.3}$$

($i = 1, \cdots, r$) are uniquely determined independently of the choice of a maximal chain [32], [33].

(3) *A partial order ($\prec$) is defined on $\mathscr{P} \setminus \{C_0, C_\infty\}$ by the relation*

$$C_i \prec C_j \quad \textit{iff } C_j \subset X \in L(p) \textit{ implies } C_i \subset X$$

where $1 \leqq i, j \leqq r$. The partial order is trivially extended over to $\mathscr{P}$ by

$$C_0 \prec C_i \prec C_\infty \quad \textit{for } i = 1, \cdots, r,$$

if C_0 and/or C_∞ are nonempty.

(4) *The "minors" defined in* (2) *above are expressed also as*

$$p_i(Y) = p(\langle C_i \rangle \cup Y) - p(\langle C_i \rangle), \qquad Y \subset C_i, \tag{2.4}$$

for $i = 1, \cdots, r$, where

$$\langle C_i \rangle = \cup \{C_j | C_j \prec C_i, C_j \neq C_i\}. \tag{2.5}$$

Note that a linear extension ($\leqq$) of the partial order defined in (3) above can be obtained by choosing a maximal chain in $L(p)$ as in (1) and by defining the total order on $\mathscr{P}$ by

$$C_i \leqq C_j \quad \text{iff } i \leqq j.$$

We write $C_i | \prec C_j$ iff $C_i \prec C_j$ and there exists no $C_k (\neq C_i, C_j)$ such that $C_i \prec C_k \prec C_j$.

3. Mixed matrices and layered mixed matrices. Let **K** be a field, which contains **Q**, the field of rationals, and of which **F** is an extension field:

$$\mathbf{Q} \subset \mathbf{K} \subset \mathbf{F}. \tag{3.1}$$

The set of $m \times n$ matrices over **F** is denoted as $\mathscr{M}(\mathbf{F}; m, n)$ or simply as $\mathscr{M}(\mathbf{F})$.

A matrix $A \in \mathscr{M}(\mathbf{F})$ can be expressed as

$$A = Q_A + T_A \tag{3.2}$$

in such a way that $Q_A \in \mathscr{M}(\mathbf{K})$ and the nonvanishing entries of T_A are in $\mathbf{F} \setminus \mathbf{K}$. To make the decomposition unique, we will assume that $(Q_A)_{ij} = 0$ if $(T_A)_{ij} \neq 0$. If, in addition, the collection of the nonvanishing entries of T_A is algebraically independent [43] over $\mathbf{K}$, the matrix A is called a *mixed matrix* with respect to $\mathbf{K}$. We denote by $\mathcal{MM}(\mathbf{F}/\mathbf{K};\, m, n)$ the set of $m \times n$ mixed matrices over $\mathbf{F}$ with respect to $\mathbf{K}$. The notion of mixed matrix is introduced in [27], [28] as a mathematical tool for dealing with structural aspects of physical/engineering systems. See [28] for detailed discussion of its physical meanings.

A subclass of mixed matrices is defined here. We call a mixed matrix $A \in \mathcal{MM}(\mathbf{F}/\mathbf{K};\, m, n)$ a *layered mixed matrix* with respect to $\mathbf{K}$, if the sets of nonzero rows of Q_A and T_A are disjoint in the expression (3.2) for a mixed matrix A, i.e., if A can be put into a partitioned matrix of the form

$$A = \begin{pmatrix} Q \\ \hline T \end{pmatrix} \tag{3.3}$$

where $Q \in \mathcal{M}(\mathbf{K};\, m_Q, n)$, $T \in \mathcal{M}(\mathbf{F};\, m_T, n)$ $(m_Q + m_T = m)$, and the collection of the nonvanishing entries of T are algebraically independent over $\mathbf{K}$. The set of $m \times n$ layered mixed matrices consisting of $m_Q + m_T$ rows as above will be designated by $\mathcal{LM}(\mathbf{F}/\mathbf{K};\, m_Q, m_T, n)$ or simply by $\mathcal{LM}(\mathbf{F}/\mathbf{K})$. Obviously we have

$$\mathcal{LM}(\mathbf{F}/\mathbf{K}; m_Q, m_T, n) \subset \mathcal{MM}(\mathbf{F}/\mathbf{K}; m_Q + m_T, n). \tag{3.4}$$

Consider a system of equations (1.1) where the coefficient matrix $A \in \mathcal{MM}(\mathbf{F}/\mathbf{K};\, m, n)$ is of the form (3.2). Introducing an auxiliary vector $\mathbf{w} \in \mathbf{R}^m$, we can express it equivalently as

$$\begin{pmatrix} I_m & Q_A \\ -I_m & T_A \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \tag{3.5}$$

It may be assumed that we can choose m numbers in $\mathbf{F}$, say $t_1, \cdots, t_m$, that are algebraically independent over the subfield of $\mathbf{F}$ to which the entries of T_A belong. Then, multiplying each of the last m equations by the transcendentals $t_1, \cdots, t_m$, we obtain an augmented system of equations

$$\begin{pmatrix} I_m & Q_A \\ -D_m & D_m T_A \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}, \tag{3.6}$$

$$D_m = \operatorname{diag}(t_1, \cdots, t_m), \tag{3.7}$$

which is still equivalent to the original system (1.1). The coefficient matrix of (3.6) is a layered mixed matrix with respect to $\mathbf{K}$ since the nonvanishing entries of $[-D_m | D_m T_A]$ are algebraically independent over $\mathbf{K}$. In the case of a system of linear/nonlinear equations (1.4), the above transformation from (1.1) to (3.5) may be interpreted as assigning w to the nonlinear part $\mathbf{g}(\mathbf{x})$ to obtain

$$\mathbf{w} + Q\mathbf{x} = \mathbf{0}, \qquad -\mathbf{w} + \mathbf{g}(\mathbf{x}) = \mathbf{0}, \tag{3.8}$$

which is equivalent to (1.4).

In general, with a mixed matrix $A \in \mathcal{MM}(\mathbf{F}/\mathbf{K};\, m, n)$ we will associate a layered mixed matrix $\tilde{A} \in \mathcal{LM}(\mathbf{F}/\mathbf{K};\, m, m, m + n)$:

$$\tilde{A} = \begin{pmatrix} I_m & Q_A \\ -D_m & D_m T_A \end{pmatrix} \tag{3.9}$$

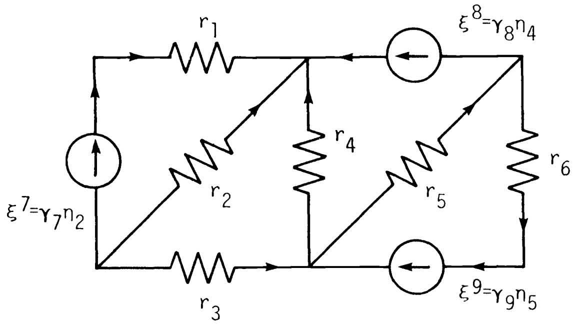


FIG. 3.1. *A simple electrical network of Example* 3.1 (*from* [30]).

where D_m is given by (3.7). Note that the column-set of $\tilde{A}$ has a natural one-to-one correspondence with the union of the column- and the row-set of A. Since we have the obvious identity

$$\text{rank}\,\tilde{A} = \text{rank}\,A + m, \tag{3.10}$$

we may restrict ourselves to layered mixed matrices when we deal with the unique solvability of a system of equations having a mixed matrix as its coefficient matrix.

For a matrix G over a field in general, we will denote by $\mathbf{M}(G)$ the linear matroid [44] defined on the column-set of G with respect to the linear dependence of the column-vectors. The rank of a layered mixed matrix A of (3.3) is known [44] (cf. also [9]) to be expressed as follows in terms of the rank of the union $\mathbf{M}(Q) \vee \mathbf{M}(T)$ of two matroids $\mathbf{M}(Q)$ and $\mathbf{M}(T)$. Both $\mathbf{M}(Q)$ and $\mathbf{M}(T)$ are defined on the column-set, say C, of the matrix A, and their rank functions will be denoted by ρ and τ, respectively.

THEOREM 3.1. *Let* $A \in \mathcal{LM}(\mathbf{F}/\mathbf{K};\, m_Q,\, m_T,\, n)$ *be a layered mixed matrix of the form* (3.3). *Then*

$$\begin{aligned}\text{rank}\,A &= \text{rank}\,(\mathbf{M}(Q) \vee \mathbf{M}(T))\\ &= \min\,\{\rho(X) + \tau(X) - |X| \,|\, X \subset C\} + n.\end{aligned}$$

Proof. By the generalized Laplace expansion and the well-known identity for matroid union. □

Note that the rank of the union of two matroids can be found by an efficient practical algorithm either for matroid union or for matroid intersection [3], [8], [20], [42].

COROLLARY 3.2 [28]. *Let* $A \in \mathcal{MM}(\mathbf{F}/\mathbf{K};\, m,\, n)$ *be a mixed matrix of the form* (3.2). *Then*

$$\text{rank}\,A = \text{rank}\,(\mathbf{M}(I_m|Q_A) \vee \mathbf{M}(I_m|T_A)) - m.$$

Proof. Immediate from (3.10) and Theorem 3.1. □

Example 3.1 [30, Example 4.1.3]. Consider the free electrical network of Fig. 3.1, which is taken from [30]. It consists of 6 resistors of resistances r_i (branch i) ($i = 1, \cdots, 6$), and 3 voltage-controlled current sources (branch i) with mutual conductances γ_i ($i = 7, 8, 9$); the current sources of branches 7, 8, 9 are controlled, respectively, by the voltages across branches 2, 4, 5. Then the current ξ^i in and the voltage η_i across branch i ($i = 1, \cdots, 9$) are to satisfy the structural equations (Kirchhoff's laws) and the constitutive equations, which altogether are expressed as in (1.1) with $\mathbf{x} = (\xi^1, \cdots, \xi^9;\, \eta_1, \cdots, \eta_9)$, $\mathbf{b} = \mathbf{0}$ and

$$(3.11)\quad A=\begin{array}{ccccccccc|ccccccccc}
\xi^1 & \xi^2 & \xi^3 & \xi^4 & \xi^5 & \xi^6 & \xi^7 & \xi^8 & \xi^9 & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5 & \eta_6 & \eta_7 & \eta_8 & \eta_9 \\
\hline
1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & & & & & & & & & \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & & & & & & & & & \\
0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & -1 & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & & & & & & & & & \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & & & & & & & & & \\
 & & & & & & & & & 0 & -1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & -1 & 0 & 1 & 1 & 0 & 0 & -1 & 0 & 0 \\
 & & & & & & & & & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 0 \\
 & & & & & & & & & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
\hdashline
r_1 & & & & & & & & & -1 & & & & & & & & \\
 & r_2 & & & & & & & & & -1 & & & & & & & \\
 & & r_3 & & & & & & & & & -1 & & & & & & \\
 & & & r_4 & & & & & & & & & -1 & & & & & \\
 & & & & r_5 & & & & & & & & & -1 & & & & \\
 & & & & & r_6 & & & & & & & & & -1 & & & \\
 & & & & & & -1 & & & & \gamma_7 & & & & & 0 & & \\
 & & & & & & & -1 & & & & & \gamma_8 & & & & 0 & \\
 & & & & & & & & -1 & & & & & \gamma_9 & & & & 0 \\
\hline
\end{array}$$

The unique solvability of the network reduces to the nonsingularity of the matrix A.

It may be justified for physical reasons (see, e.g., [28]) to regard r_i $(i = 1, \cdots, 6)$ and γ_i $(i = 7, 8, 9)$ as real numbers which are collectively algebraically independent over the field of rationals. Then we have $A \in \mathcal{M}\mathcal{M}(\mathbf{R}/\mathbf{Q};\ 18, 18)$, and the unique solvability of the network can be determined efficiently by Corollary 3.2. Or alternatively [17], [31], [32], [37], [38], [39], we may directly apply Theorem 3.1 with Q being the upper 9 rows of A and T being the lower 9 rows of A, since we can put A in the form of a layered mixed matrix by multiplying the lower 9 rows by independent transcendentals, just as we did for (3.5) to get (3.6). This example will be taken up again in Example 4.2.

4. Combinatorial canonical form of layered mixed matrices. This section defines a block-triangular canonical form for an $m \times n$ layered mixed matrix $A \in \mathcal{L}\mathcal{M}(\mathbf{F}/\mathbf{K};\ m_Q, m_T, n)$ of the form (3.3), where $m = m_Q + m_T$. For A of (3.3), we consider the transformation of the form

$$(4.1)\qquad P_r\begin{pmatrix} S_Q & 0 \\ 0 & P_T \end{pmatrix}\begin{pmatrix} Q \\ T \end{pmatrix}P_c$$

where S_Q is an $m_Q \times m_Q$ nonsingular matrix over $\mathbf{K}$ (i.e., $S_Q \in GL(m_Q, \mathbf{K})$); P_T, P_r and P_c are permutation matrices of orders m_T, m and n, respectively. The transformed matrix of (4.1) also belongs to $\mathcal{L}\mathcal{M}(\mathbf{F}/\mathbf{K};\ m_Q,\ m_T,\ n)$ and is equivalent to A in the ordinary sense in linear algebra. We will say that two matrices are *LM-equivalent* if they are connected by the transformation above. In the following, we will look for a canonical block-triangular matrix among the matrices LM-equivalent to A. The canonical form to be considered should reduce to the DM-decomposition when $m = m_T$ and $m_Q = 0$.

Let R and C denote the row- and the column-set of A, respectively; the former is the disjoint union of the row-sets, say R_Q and R_T, of Q and T:

$$(4.2)\qquad R = R_Q \cup R_T.$$

For $I \subset R$ and $J \subset C$, $A[I, J]$ means the submatrix of A with row-set I and column-set J.

Theorem 3.1 states that the rank of $A[R, J]$ ($J \subset C$) can be expressed by $\rho(X) = \operatorname{rank} Q[R_Q, X]$ and $\tau(X) = \operatorname{rank} T[R_T, X]$ ($X \subset J$). On account of the algebraic independence of the nonvanishing entries of T, the rank $\tau(X)$ equals the term-rank [34] of $T[R_T, X]$, which is known [44] to be expressed by the adjacency associated with T; namely we have

$$\tau(X) = \min \{\gamma(Y) + |X \backslash Y| \mid Y \subset X\}, \qquad X \subset C, \tag{4.3}$$

where

$$\Gamma_T(Y) = \{i \in R_T \mid T_{ij} \neq 0 \text{ for some } j \in Y\}, \qquad Y \subset C, \tag{4.4}$$

$$\gamma(Y) = |\Gamma_T(Y)|, \qquad Y \subset C. \tag{4.5}$$

We consider two functions:

$$p_\tau(X) = \rho(X) + \tau(X) - |X|, \qquad X \subset C, \tag{4.6}$$

$$p_\gamma(X) = \rho(X) + \gamma(X) - |X|, \qquad X \subset C. \tag{4.7}$$

Since $\tau(X) \leqq \gamma(X)$ by definition, we have the obvious inequality

$$p_\tau(X) \leqq p_\gamma(X). \tag{4.8}$$

These functions, however, share a common minimum value when restricted to 2^J for any $J \subset C$.

LEMMA 4.1. *For $J \subset C$, we have*

$$\min \{p_\tau(X) \mid X \subset J\} = \min \{p_\gamma(X) \mid X \subset J\}.$$

Proof. From (4.6) and (4.3) it follows that

$$\begin{aligned} &\min \{p_\tau(X) \mid X \subset J\} \\ &\quad = \min \{\rho(X) - |X| + \min \{\gamma(Y) + |X \backslash Y| \mid Y \subset X\} \mid X \subset J\} \\ &\quad = \min \{\rho(X) + \gamma(Y) - |Y| \mid Y \subset X \subset J\} \\ &\quad = \min \{\rho(Y) + \gamma(Y) - |Y| \mid Y \subset J\} \\ &\quad = \min \{p_\gamma(Y) \mid Y \subset J\}. \end{aligned}$$

Combined with Theorem 3.1, this gives a characterization of rank A in terms of ρ and γ, instead of ρ and τ.

THEOREM 4.2 [25]. *Let $A \in \mathscr{LM}(\mathbf{F}/\mathbf{K}; m_Q, m_T, n)$ be of the form* (3.3). *Then*

$$\operatorname{rank} A[R, J] = \min \{p_\gamma(X) \mid X \subset J\} + |J|,$$

for $J \subset C$, where p_γ is defined by (4.7).

The important fact is that $p_\gamma : 2^C \rightarrow \mathbf{R}$ of (4.7) is submodular, and hence, as explained in §2, its minimizer $L(p_\gamma)$ determines a unique partition of the column-set C of A into partially ordered blocks. To be specific, we choose (cf. Theorem 2.1 (1)) a maximal chain in $L(p_\gamma)$:

$$X_0 \subsetneqq X_1 \subsetneqq \cdots \subsetneqq X_r \tag{4.9}$$

to get the blocks:

$$C_0 = X_0; \quad C_j = X_j \backslash X_{j-1} \quad (j = 1, \cdots, r); \quad C_\infty = C \backslash X_r. \tag{4.10}$$

We define $C_0 \prec C_j$ (resp. $C_j \prec C_\infty$) for $j = 1, \cdots, r$ if C_0 (resp. C_∞) is nonempty.

A partition $\{R_{Tj} | j = 0, 1, \cdots, r, \infty\}$ of the row-set R_T of T is induced from (4.9) naturally as follows:

$$(4.11) \qquad R_{T0} = Y_{T0}; \quad R_{Tj} = Y_{Tj} \backslash Y_{T,j-1} \qquad (j = 1, \cdots, r); \quad R_{T\infty} = R_T \backslash Y_{Tr}$$

where

$$(4.12) \qquad Y_{Tj} = \Gamma_T(X_j) \qquad (j = 0, 1, \cdots, r).$$

By this construction, we have $T[R_{Ti}, C_j] = 0$ for $i > j$, i.e., the matrix T is already essentially "block-triangularized" with respect to the partitions (4.10) and (4.11). Introducing permutation matrices P_c and P_T, we can make $\bar{T} = P_T T P_c$ in an explicit block-triangular form in the ordinary sense, where, however, the column-sets (resp. row-sets) of T and $\bar{T}$ are identified with each other by the one-to-one correspondence through the permutation P_c (resp. P_T), so that $\bar{T}[R_{Ti}, C_j] = T[R_{Ti}, C_j]$ $(0 \leqq i, j \leqq \infty)$. To be more precise, we have the following.

LEMMA 4.3.

$$R_{Tj} = \Gamma_T(\langle C_j \rangle \cup C_j) \backslash \Gamma_T(\langle C_j \rangle)$$
$$= \Gamma_T(C_j) \backslash \Gamma_T(\langle C_j \rangle) \qquad (j = 1, \cdots, r)$$

where $\langle C_j \rangle$ is defined by (2.5), *and therefore*

$$\bar{T}[R_{Ti}, C_j] = 0 \quad \textit{unless } C_i \prec C_j.$$

Proof. Since X_{j-1}, $X_j (= X_{j-1} \cup C_j)$, $\langle C_j \rangle$ and $\langle C_j \rangle \cup C_j$ all belong to $L(p_\gamma)$, we have $p_\gamma(X_{j-1} \cup C_j) - p_\gamma(X_{j-1}) = p_\gamma(\langle C_j \rangle \cup C_j) - p_\gamma(\langle C_j \rangle)$ $(=0)$. This implies, by submodularity, that

$$(4.13) \qquad \rho(X_{j-1} \cup C_j) - \rho(X_{j-1}) = \rho(\langle C_j \rangle \cup C_j) - \rho(\langle C_j \rangle),$$

$$(4.14) \qquad \gamma(X_{j-1} \cup C_j) - \gamma(X_{j-1}) = \gamma(\langle C_j \rangle \cup C_j) - \gamma(\langle C_j \rangle).$$

The latter is equivalent to $|\Gamma_T(C_j) \backslash \Gamma_T(X_{j-1})| = |\Gamma_T(C_j) \backslash \Gamma_T(\langle C_j \rangle)|$, which means $R_{Tj} = \Gamma_T(C_j) \backslash \Gamma_T(\langle C_j \rangle)$ since $R_{Tj} = \Gamma_T(C_j) \backslash \Gamma_T(X_{j-1})$ and $\Gamma_T(X_{j-1}) \supset \Gamma_T(\langle C_j \rangle)$. □

As for the matrix Q, it can be transformed to a block-triangular matrix $\bar{Q}$ with respect to the partition (4.10) by the usual elimination operations; that is, for some $S_Q \in GL(m_Q, \mathbf{K})$, the row-set of $\bar{Q} = S_Q Q P_c$ is partitioned into disjoint subsets $\{R_{Qj} | j = 0, 1, \cdots, r, \infty\}$ such that

$$|R_{Q0}| = \rho(X_0),$$
$$(4.15) \qquad |R_{Qj}| = \rho(X_j) - \rho(X_{j-1}) \qquad (j = 1, \cdots, r),$$
$$|R_{Q\infty}| = |R_Q| - \rho(X_r),$$

and

$$(4.16) \qquad \bar{Q}[R_{Qi}, C_j] = 0 \qquad (0 \leqq j < i \leqq \infty).$$

By the same argument as the proof of Lemma 4.3 (by (4.13) in particular), we see

$$|R_{Qj}| = \rho(\langle C_j \rangle \cup C_j) - \rho(\langle C_j \rangle)$$

and we may further assume that

$$(4.17) \qquad \bar{Q}[R_{Qi}, C_j] = 0 \quad \text{unless } C_i \prec C_j.$$

We will put

$$Y_{Qj} = \bigcup_{i=0}^{j} R_{Qi} \qquad (j=0,1,\cdots,r), \tag{4.18}$$

$$Y_j = Y_{Qj} \cup Y_{Tj} \qquad (j=0,1,\cdots,r), \tag{4.19}$$

$$R_j = R_{Qj} \cup R_{Tj} \qquad (j=0,1,\cdots,r,\infty). \tag{4.20}$$

It may be noted that, if we require (4.16) only (not necessarily (4.17)), we can choose S_Q to be expressed as

$$S_Q = L_Q P_Q \tag{4.21}$$

where $L_Q \in GL(m_Q, \mathbf{K})$ is lower block-triangular and P_Q is a permutation matrix.

Consider the matrix

$$\bar{A} = \begin{pmatrix} \bar{Q} \\ \bar{T} \end{pmatrix} = \begin{pmatrix} S_Q Q P_c \\ P_T T P_c \end{pmatrix}, \tag{4.22}$$

which is LM-equivalent to A (under the transformation (4.1)). The row-set $R = R_Q \cup R_T$ of $\bar{A}$, as well as the column-set C, is now partitioned into blocks $\{R_j | j = 0, 1, \cdots, r, \infty\}$, on which the partial order $(\prec)$ on $\{C_j | j = 0, 1, \cdots, r, \infty\}$ can naturally be induced.

THEOREM 4.4. *Let $\bar{A}$ be as above, whose row-set R and column-set C are partitioned into partially ordered blocks.*

(1) $\bar{A}[R_i, C_j] = 0$ *unless* $C_i \prec C_j$ $(1 \leqq i, j \leqq r)$. *In particular,*

$$\bar{A}[R_i, C_j] = 0 \quad \textit{if}\ i > j. \tag{4.23}$$

$\bar{A}[R_i, C_j] \neq 0$ *if* $C_i | \!\prec C_j$ $\quad (1 \leqq i, j \leqq r)$.

(2) $|R_0| < |C_0|$ *if* $C_0 \neq \emptyset$,
$|R_j| = |C_j|\ (>0)$ *for* $j = 1, \cdots, r$,
$|R_\infty| > |C_\infty|$ *if* $C_\infty \neq \emptyset$.
(*From the last relation follows a more symmetric but weaker statement*:
$|R_\infty| \underline{>} |C_\infty|$ *if* $R_\infty \neq \emptyset$.)

(3) rank $\bar{A}[Y_j, X_j]$ = rank $\bar{A}[R, X_j] = |Y_j|$ $\quad (j = 0, 1, \cdots, r)$.

(4) rank $\bar{Q}[Y_{Qj}, X_j] = |Y_{Qj}|$ $\quad (j = 0, 1, \cdots, r)$,
rank $\bar{T}[Y_{Tj}, X_j] = |Y_{Tj}|$ $\quad (j = 0, 1, \cdots, r)$.

(5) rank $\bar{A}[R_0, C_0] = |R_0|$,
rank $\bar{A}[R_j, C_j] = |R_j| = |C_j|\ (>0)$ $\quad (j = 1, \cdots, r)$,
rank $\bar{A}[R_\infty, C_\infty] = |C_\infty|$.

(6) *For* $j = 0, 1, \cdots, r, \infty$, *the submatrix* $\bar{A}[R_j, C_j]$ $(\in \mathscr{LM}(\mathrm{F}/\mathrm{K}))$ *is irreducible in the sense that the submodular function* $\bar{p}_j$ (*defined on* C_j), *the correspondent of* p_γ *of* (4.7), *has no minimizers distinct from* $\emptyset$ *and* C_j.

Proof. (1): Immediate from Lemma 4.3 and (4.17).

(2): If $C_0 \neq \emptyset$, then $0 = p_\gamma(\emptyset) > \min p_\gamma = p_\gamma(C_0) = \rho(C_0) + \gamma(C_0) - |C_0| = |R_0| - |C_0|$.

For $j = 1, \cdots, r$, we have $p_\gamma(X_{j-1}) = p_\gamma(X_j)$, i.e.,

$$\rho(X_{j-1}) + \gamma(X_{j-1}) - |X_{j-1}| = \rho(X_j) + \gamma(X_j) - |X_j|.$$

By (4.11), (4.12), and (4.15), this reduces to $|C_j| = |R_j|$.

If $C_\infty \neq \emptyset$, then $p_\gamma(C) > \min p_\gamma = p_\gamma(X_r)$, which implies $|R| - |C| \geqq \rho(C) + \gamma(C) - |C| > \rho(X_r) + \gamma(X_r) - |X_r| = |Y_r| - |X_r|$. Hence $|R_\infty| = |R| - |Y_r| > |C| - |X_r| = |C_\infty|$.

(3): From (1) above and Theorem 4.2, we have rank $\bar{A}[Y_j, X_j]$ = rank $\bar{A}[R, X_j]$ = rank $A[R, X_j]$ = min $\{p_\gamma(X) | X \subset X_j\}$ + $|X_j|$ = $p_\gamma(X_j)$ + $|X_j|$ = $\rho(X_j)$ + $\gamma(X_j)$ = $|Y_{Qj}|$ + $|Y_{Tj}|$ = $|Y_j|$.

(4): Immediately from (3) above.

(5): The identities for $j = 0, 1, \cdots, r$ are immediate from (1) and (3) above. By Theorem 4.2, we have

$$\text{rank}\, \bar{A}[R_\infty, C_\infty] = \min\, \{\bar{p}_\infty(Z) | Z \subset C_\infty\} + |C_\infty|$$

where

$$\bar{p}_\infty(Z) = \text{rank}\, \bar{Q}[R_{Q\infty}, Z] + |\Gamma_T(Z) \cap R_{T\infty}| - |Z|.$$

On the other hand, this turns out to be nonnegative for $Z \subset C_\infty$, since

$$\begin{aligned} \bar{p}_\infty(Z) &= (\rho(X_r \cup Z) - \rho(X_r)) + (\gamma(X_r \cup Z) - \gamma(X_r)) - |Z| \\ &= p_\gamma(X_r \cup Z) - p_\gamma(X_r) \qquad (4.24) \\ &= p_\gamma(X_r \cup Z) - \min p_\gamma. \end{aligned}$$

(6): First consider the case of $j = \infty$. Recalling $X_r = \max L(p_\gamma)$, we see from (4.24) that $\bar{p}_\infty$ has the unique minimizer $Z = \varnothing$. The second case of $j = 0$ is easy, since $\bar{p}_0(Z) = p_\gamma(Z)$ has the unique minimizer $Z = C_0$. The other cases $(1 \leqq j \leqq r)$ can be treated similarly using the expression

$$\begin{aligned} \bar{p}_j(Z) &= \text{rank}\, \bar{Q}[R_{Qj}, Z] + |\Gamma_T(Z) \cap R_{Tj}| - |Z| \\ &= p_\gamma(X_{j-1} \cup Z) - \min p_\gamma. \end{aligned}$$ □

This theorem shows that with suitable permutation matrix P_r, $P_r\bar{A}$ is a block-triangular matrix which is LM-equivalent to A. The ordering of the blocks is uniquely determined in the sense of the partial order ($\prec$). The following states that it is the finest block-triangular form that is LM-equivalent to A.

THEOREM 4.5. *The matrix $P_r\bar{A}$ constructed above based on p_γ is the finest block-triangular matrix that is LM-equivalent to A and enjoys the properties* (2) *and* (5) *of Theorem* 4.4.

Proof. Suppose that $\hat{A}$ is such a block-triangular matrix with the row-set R and the column-set C being partitioned as

$$R = \cup\{R'_j | j = 0, 1, \cdots, r', \infty\}, \qquad C = \cup\{C'_j | j = 0, 1, \cdots, r', \infty\}, \tag{4.25}$$

where $\hat{A}[R'_i, C'_j] = 0$ for $i > j$. Since $\hat{A}$ is LM-equivalent to A, we have from Theorem 4.2

$$\text{rank}\, \hat{A} = \min\, \{p_\gamma(X) | X \subset C\} + |C| \tag{4.26}$$

with the same p_γ as for A. Put

$$X'_j = \bigcup_{i=0}^{j} C'_i \qquad (j = 0, 1, \cdots, r'), \tag{4.27}$$

$$Y'_j = \bigcup_{i=0}^{j} R'_i \qquad (j = 0, 1, \cdots, r'). \tag{4.28}$$

Since $\hat{A}$ is block-triangularized and has the property (5) of Theorem 4.4, we have

$$\text{rank}\, \hat{A} = |C| - |X'_j| + |Y'_j| \qquad (j = 0, 1, \cdots, r'). \tag{4.29}$$

Combining (4.26) and (4.29), we obtain

$$\min p_\gamma = |Y'_j| - |X'_j| \qquad (j = 0, 1, \cdots, r').$$

This shows that

$$X_j' \in L(p_\gamma), \tag{4.30}$$

since $p_\gamma(X_j') = \rho(X_j') + \gamma(X_j') - |X_j'| \leqq |Y_j'| - |X_j'| = \min p_\gamma$. Therefore, the partition (4.25) is coarser than or equal to (or an aggregation of) $\{C_j | j = 0, 1, \cdots, r, \infty\}$ determined by $L(p_\gamma)$. □

Thus, the matrix $P_r\bar{A}$ with $\bar{A}$ constructed above provides the finest block-triangular form among the matrices LM-equivalent to A. It is named here the *combinatorial canonical form of a layered mixed matrix*. It is obvious that it agrees with the DM-decomposition when $A = T$ (i.e., $m_Q = 0$). In parallel with the DM-decomposition, the rectangular blocks corresponding to $R_0 \times C_0$ and $R_\infty \times C_\infty$, if any, will be called the *horizontal tail* and the *vertical tail*, respectively.

A comment on the algorithm will be in order. From the point of view of practical application, it is important to note that this canonical form can be constructed by an efficient matroid-theoretic algorithm that involves $O(n^3 \log n)$ arithmetic operations [3] in the subfield **K** and $O((m + n)^2 n)$ operations for graph manipulations, as follows.

To be specific, with $A \in \mathscr{LM}(\mathbf{F}/\mathbf{K};\ m_Q,\ m_T,\ n)$ having the row-set $R = R_Q \cup R_T$ and the column-set C we associate a bipartite graph $G = (R_T \cup C_Q, C; E)$ defined as follows. The vertex-set V of G is given by

$$V = R_T \cup C_Q \cup C \tag{4.31}$$

where C_Q is a disjoint copy of C, and the arc-set E of G is defined as

$$E = \{(i,j) \in R_T \times C | T_{ij} \neq 0\} \cup \{(j_Q, j) \in C_Q \times C | j \in C\} \tag{4.32}$$

where j_Q $(\in C_Q)$ denotes the copy of j $(\in C)$.

We consider the independent-flow problem [10] (see also [20]) on the network with the underlying graph G (or an independent-matching problem [44]); the direct sum of a free matroid on R_T and the linear matroid $\mathbf{M}(Q)$ on C_Q is defined on the entrance-set $R_T \cup C_Q$, another free matroid is attached to the exit-set C, and each arc of E has infinite capacity. For $U \subset V$ we put

$$J = C \backslash U, \quad I = R_T \backslash U, \quad K_Q = C_Q \backslash U. \tag{4.33}$$

Then the capacity $\kappa(U)$ of U is given by

$$\kappa(U) = \begin{cases} |I| + \rho(K) + |C \backslash J| & \text{if } \Gamma_T(J) \subset I \text{ and } J \subset K, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.34}$$

where K_Q $(\subset C_Q)$ and K $(\subset C)$ are corresponding copies. The family $L(\kappa)$ of the minimizers of κ, namely the family of minimum cuts, determines $L(p_\gamma)$ by

$$L(p_\gamma) = \{J \subset C | J = C \backslash U, U \in L(\kappa)\}. \tag{4.35}$$

This shows that the desired partition of C for the combinatorial canonical form can be constructed by first finding the maximum independent flow (or independent matching) and then decomposing the auxiliary graph associated with it into strongly connected components, among which the partial order can be induced. (To be more precise, the column-sets C_0 and C_∞ are determined by those vertices of C $(\subset V)$ which are reachable to the exit and from the entrance, respectively.) See, e.g., [20] for detail. Example 4.1 below will illustrate this procedure.

Example 4.1. Consider the following matrix $A \in \mathscr{LM}(\mathbf{F}/\mathbf{Q};\ 3,\ 6,\ 7)$, where $\{t_i | i = 1, \cdots, 13\}$ are indeterminates over **Q** and **F** is the field of rational functions in t_i's over **Q**:

$$
(4.36) \qquad A = \begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & 0 & 1 & 0 & 1 & -1 \\
-2 & 0 & 1 & -2 & 0 & 0 & 2 \\
1 & 0 & 0 & 1 & 1 & 1 & -1 \\
\hdashline
t_1 & & & & & t_2 & \\
t_3 & & & & t_4 & & \\
 & t_5 & & t_6 & t_7 & & \\
 & t_8 & & t_9 & t_{10} & & t_{11} \\
 & & & & & t_{12} & \\
 & & & & & t_{13} & \\
\hline
\end{array} \; .
$$

The graph G for the associated independent-flow problem is depicted in Fig. 4.1. The auxiliary graph for a maximum independent flow is shown in Fig. 4.2, which provides the partition (4.10) of the column-set C of A:

$$
(4.37) \qquad C = C_0 \cup C_1 \cup C_2 \cup C_\infty
$$

where $C_0 = \emptyset$, $C_1 = \{2, 4, 7\}$, $C_2 = \{3\}$, $C_\infty = \{1, 5, 6\}$; C_0 (resp. C_∞) consists of those vertices of C which are reachable to s^- (resp. from s^+), and C_1 and C_2 are determined by the strong components of the subgraph of the auxiliary graph that is obtained by deleting the vertices reachable to s^- or from s^+. Notice $C_i \prec C_\infty$ $(i = 1, 2)$, and that C_1 and C_2 have no order relation with each other. The combinatorial canonical form of A is given by

$$
(4.38) \qquad \begin{array}{ccc|c|ccc}
2 & 4 & 7 & 3 & 1 & 5 & 6 \\
\hline
0 & 1 & -1 & & 1 & & 1 \\
t_5 & t_6 & 0 & & & t_7 & \\
t_8 & t_9 & t_{11} & & & t_{10} & \\
\hline
 & & & \boxed{1} & & & 2 \\
\hline
 & & & & 0 & 1 & 0 \\
 & & & & t_1 & 0 & t_2 \\
 & & & & t_3 & t_4 & 0 \\
 & & & & 0 & 0 & t_{12} \\
 & & & & 0 & 0 & t_{13} \\
\hline
\end{array} \; .
$$

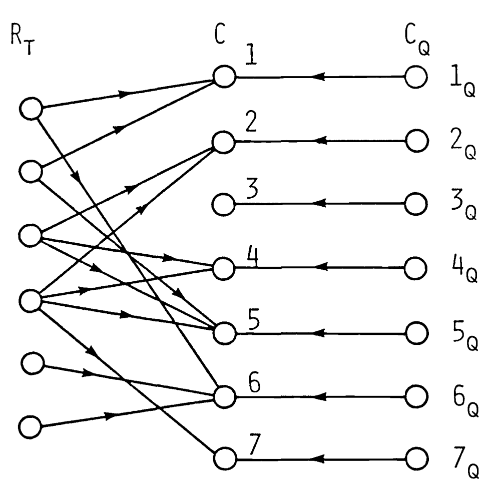


FIG. 4.1. *Independent-flow problem for Example* 4.1.

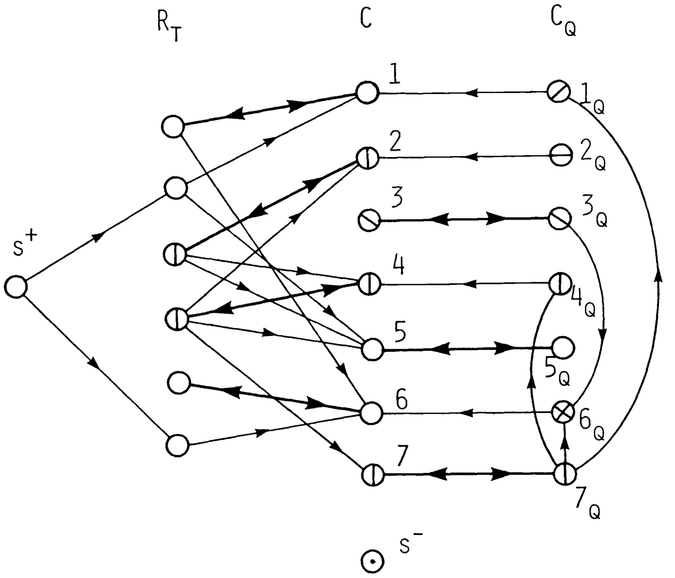


FIG. 4.2. *Auxiliary graph associated with a maximal independent flow for Example* 4.1.

Example 4.2. Recall the electrical network of Example 3.1. With the understanding, mentioned in Example 3.1, that the coefficient matrix A of (3.11) can be considered a member of $\mathscr{LM}(\mathbf{R}/\mathbf{Q}; 9, 9, 18)$, the combinatorial canonical form of A is found as (4.39) below.

It has empty tails ($C_0 = R_\infty = \emptyset$) and 9 square diagonal blocks with the column-sets given by $C_1 = \{\eta_7\}$, $C_2 = \{\eta_1\}$, $C_3 = \{\xi^1\}$, $C_4 = \{\eta_8\}$, $C_5 = \{\eta_9\}$, $C_6 = \{\eta_6\}$, $C_7 = \{\xi^6\}$, $C_8 = \{\eta_5, \xi^5, \xi^9\}$, $C_9 = \{\xi^2, \eta_2, \xi^3, \eta_3, \xi^4, \eta_4, \xi^7, \xi^8\}$. The partial order among them is given by:

$$C_1 \prec C_2 \prec C_3 \prec C_9; C_4 \prec C_8 \prec C_9; C_5 \prec C_6 \prec C_7 \prec C_8.$$

$$\begin{array}{c|c|c|c|c|c|c|ccc|cccccccc}
\eta_7 & \eta_1 & \xi^1 & \eta_8 & \eta_9 & \eta_6 & \xi^6 & \eta_5 & \xi^5 & \xi^9 & \xi^2 & \eta_2 & \xi^3 & \eta_3 & \xi^4 & \eta_4 & \xi^7 & \xi^8 \\
\hline
-1 & -1 & & & & & & & & & & & & 1 & & 1 & & \\
 & -1 & r_1 & & & & & & & & & & & & & & & \\
 & & 1 & & & & & & & & & & & & & & -1 & \\
 & & & -1 & & & & -1 & & & & & & & & 1 & & \\
 & & & & -1 & -1 & & -1 & & & & & & & & & & \\
 & & & & & -1 & r_6 & & & & & & & & & & & \\
 & & & & & & 1 & & & -1 & & & & & & & & \\
 & & & & & & & 0 & 1 & -1 & 1 & & & & 1 & & 1 & \\
 & & & & & & & -1 & r_5 & 0 & & & & & & & & \\
 & & & & & & & \gamma_9 & 0 & -1 & & & & & & & & \\
\hline
 & & & & & & & & & & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
 & & & & & & & & & & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
 & & & & & & & & & & 0 & -1 & 0 & 1 & 0 & 1 & 0 & 0 \\
 & & & & & & & & & & r_2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 0 & 0 & r_3 & -1 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & r_4 & -1 & 0 & 0 \\
 & & & & & & & & & & 0 & \gamma_7 & 0 & 0 & 0 & 0 & -1 & 0 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & \gamma_8 & 0 & -1
\end{array} \tag{4.39}$$

This example will be considered again in Example 5.2.

We now consider how the combinatorial canonical form can be applied to an efficient solution of a system of equations $A(\theta)\mathbf{x} = \mathbf{b}(\theta)$ for varying values of parameters θ. We express the coefficient matrix as

$$A(\theta) = Q_A + T_A(\theta) \tag{4.40}$$

and regard it as a mixed matrix, treating the nonvanishing entries of $T_A(\theta)$ as if they were algebraically independent. As discussed at the beginning of §3, we may introduce an auxiliary variable $\mathbf{w}$ to obtain the augmented system of equations (3.5) or (3.6) with the layered mixed matrix $\tilde{A}$ of (3.9) as the coefficient matrix. The combinatorial canonical form of $\tilde{A}$ determines a hierarchical decomposition of the whole augmented system into smaller subsystems; we may repeatedly solve the subproblems with the diagonal blocks as the coefficient matrices.

For the subproblems to be solved, the diagonal blocks of the combinatorial canonical form of $\tilde{A}$ must be nonsingular. If the assumption of the algebraic independence of the nonvanishing entries of $T_A(\theta)$ is literally met, the nonsingularity of the diagonal blocks is guaranteed by Theorem 4.4(5). It is obvious, however, from the block-triangular structure that even if the assumption is not satisfied, the diagonal blocks must be nonsingular if the original coefficient matrix A is nonsingular at all. Therefore the decomposition procedure above can be carried out successfully if the original system is uniquely solvable at all.

Each subproblem may be solved as follows. Let $\bar{A}_j$ be the coefficient matrix of the jth subproblem. Its row-set is divided as (4.20) into R_{Qj} and R_{Tj}. Its column-set C_j may also be partitioned as

$$C_j = C_{wj} \cup C_{xj} \tag{4.41}$$

where C_{wj} and C_{xj} correspond to part of the variables $\mathbf{w}$ and $\mathbf{x}$, respectively. It is easy to see, by the irreducibility of $\bar{A}_j$, that

$$|R_{Tj}| \geqq |C_{wj}| \quad \text{if } R_{Tj} \neq \emptyset \tag{4.42}$$

(and $|C_j| = 1$ if $R_{Tj} = \emptyset$) and that the submatrix $\bar{A}_j[R_{Tj}, C_{wj}]$ is of the simple form

$$\bar{A}_j[R_{Tj}, C_{wj}] = \begin{pmatrix} -I \\ 0 \end{pmatrix} \tag{4.43}$$

if $R_{Tj} \neq \emptyset$ and $C_{wj} \neq \emptyset$, where I is the identity matrix of order $|C_{wj}|$. Thus the subproblem can be expressed as

$$\begin{array}{r c} & \begin{array}{cc} C_{wj} & C_{xj} \end{array} \\ \begin{array}{r} R_{Qj}\colon \\ \\ R_{Tj}\colon \end{array} & \left(\begin{array}{cc} Q_1 & Q_2 \\ \hline -I & T_1 \\ 0 & T_2 \end{array}\right) \end{array} \begin{pmatrix} \mathbf{w}_j \\ \mathbf{x}_j \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{b}}_j \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \tag{4.44}$$

where $\bar{\mathbf{b}}_j = \bar{\mathbf{b}}_j(\theta)$ is to be computed from $\mathbf{b}(\theta)$ each time θ is given. On eliminating the auxiliary variables $\mathbf{w}_j$, we obtain the system of equations

$$\begin{pmatrix} Q_1 T_1 + Q_2 \\ T_2 \end{pmatrix} \mathbf{x}_j = \begin{pmatrix} \bar{\mathbf{b}}_j \\ \mathbf{0} \end{pmatrix} \tag{4.45}$$

in $|C_{xj}|$ variables. The amount of computation needed to determine $\mathbf{x}_j$ in this way may be estimated roughly by

$$|R_{Qj}||C_{wj}||C_{xj}| + |C_{xj}|^3/3. \tag{4.46}$$

Another approach may be conceivable that makes no distinction between $\mathbf{w}_j$ and $\mathbf{x}_j$. We may assume that the subsystem is given by

$$(4.47)\qquad \begin{array}{r} R_{Qj}: \\ R_{Tj}: \end{array}\begin{pmatrix} I & Q_1 \\ T_1 & T_2 \end{pmatrix}\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{b}}_j \\ \mathbf{0} \end{pmatrix}$$

where $(\mathbf{z}_1, \mathbf{z}_2)$ is a rearrangement of $(\mathbf{w}_j, \mathbf{x}_j)$. The Gaussian elimination procedure applied to (4.46), possibly with permutations of rows in R_{Tj}, can be done with at most

$$(4.48)\qquad |R_{Tj}|^2|R_{Qj}| + |R_{Tj}|^3/3$$

arithmetic operations.

The above considerations reveal that the matrix $\bar{A}_j$ contains an identity matrix of order no smaller than max $(|C_{wj}|, |R_{Qj}|)$ as a submatrix. Thus, we may adopt

$$(4.49)\qquad \min\,(|C_{xj}|, |R_{Tj}|)$$

as a rough measure for the substantial size of the subproblem.

Example 4.3. This example is based on the reactor-separator model (EV-6) of [45]. The system of linear/nonlinear equations to be solved involves 120 unknowns and as many equations. The Jacobian matrix, denoted as A, is sparse, containing 351 nonvanishing entries. The ordinary DM-decomposition yields 4 nontrivial blocks involving more than one unknown variable. The maximum size of the blocks is 25 (see Table 4.1).

Of the nonvanishing entries of A, 172 numbers are rational constants (1 or -1) and the remaining 179 entries are regarded here as algebraically independent numbers (in a field $\mathbf{F}$) over $\mathbf{Q}$. That is, we consider $A \in \mathcal{M}\mathcal{M}(\mathbf{F}/\mathbf{Q};\ 120, 120)$. As explained above, we may then resort to the combinatorial canonical form of the corresponding layered mixed matrix $\tilde{A} \in \mathcal{L}\mathcal{M}(\mathbf{F}/\mathbf{Q};\ 120, 120, 240)$ to obtain a decomposition of the augmented system of equations with auxiliary variables (see (3.2) and (3.9)). The canonical form of $\tilde{A}$ has empty tails and yields 5 nontrivial blocks, the maximum size of which being equal to 17. (The canonical form of $\tilde{A}$ has been found by a slightly modified version of the FORTRAN program originally coded by M. Ichikawa [12].) In Table 4.1, three different decompositions are compared, where the number of rows of the T-part of each block, i.e., $|R_{Tj}|$ of (4.11), is indicated in brackets. The third decomposition will be explained in §5.

Example 4.4. The system of equations considered here is compiled in [12] from a real-world problem that has arisen from the analysis of an industrial hydrogen production system. It involves 544 variables and equations, and the Jacobian matrix A consists of 1142 rational constants (1 or -1) and 322 other numbers which are regarded here as algebraically independent transcendentals in $\mathbf{F}$ over $\mathbf{Q}$. Then we have $A \in \mathcal{M}\mathcal{M}(\mathbf{F}/\mathbf{Q};\ 544, 544)$. The combinatorial canonical form of the corresponding layered mixed matrix $\tilde{A} \in \mathcal{L}\mathcal{M}(\mathbf{F}/\mathbf{Q};\ 544, 544, 1088)$, computed as in Example 4.3, has empty

TABLE 4.1
Block-triangularizations for Example 4.3.

DM-decomposition of A		Combin. canon. form of $\tilde{A}$ (by p_γ)		Decomposition of $\tilde{A}$ by p_τ	
size	blocks	size	blocks	size	blocks
C_x		$C = C_w + C_x\ [R_T]$		$C = C_w + C_x\ [R_T]$	
25	1	17 = 8 + 9 [9]	1	16 = 8 + 8 [8]	1
10	1	15 = 6 + 9 [6]	1	14 = 6 + 8 [5]	1
9	2	14 = 4 + 10 [9]	1	13 = 4 + 9 [8]	1
		8 = 0 + 8 [4]	1	8 = 0 + 8 [5]	1
		5 = 0 + 5 [5]	1		
1	67	1	181	1	189

tails and contains 23 nontrivial blocks with more than one variable. The DM-decomposition of A and the combinatorial canonical form of $\tilde{A}$ are summarized in Table 4.2. Note that the substantial sizes of the subproblems in terms of (4.49) are much smaller than the block sizes of the subproblems obtained by the DM-decomposition.

5. Relations to other decompositions. The first subsection clarifies the relation of the combinatorial canonical form to the decomposition considered in [30], [31], [32], as well as to the ordinary DM-decomposition. The second subsection points out that for a certain class of electrical networks considered in [16], [17], [42], the combinatorial canonical form gives essentially the same block-triangularization as the method proposed in [16], [17] by way of the structure of minimum covers in an independent-matching problem.

5.1. Decomposition by $L(p_\tau)$ and the DM-decomposition. It has been claimed in [30], [31], [32] that a block-triangularization of systems of equations, such as (3.11), for electrical networks is obtained by the principal partition associated with a matroid intersection problem. The method of [30], [31], [32], which we term here the *principal partition of* $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$, is based on Theorem 3.1 and adopts the submodular function p_τ of (4.6) to obtain a decomposition of unknown variables (i.e., currents and voltages of branches in the case of electrical networks) into partially ordered blocks; that is, the principal partition of $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$ for a layered mixed matrix (3.3) is the partition of the column-set into partially ordered blocks produced by the lattice $L(p_\tau)$ (the family of minimizers of p_τ) according to Theorem 2.1. This method, however, provides too fine a partition for a block-triangularization, as is demonstrated below (see also Example 5.2).

Example 5.1. Consider an electrical network consisting of two separate branches with mutual coupling given in terms of admittances. This network is described by the matrix (cf. (5.5)):

$$A=\begin{array}{cccc} \xi^1 & \xi^2 & \eta_1 & \eta_2 \\ \hline 1 & & & \\ & 1 & & \\ \hdashline -1 & & y^{11} & y^{12} \\ & -1 & y^{21} & y^{22} \\ \hline \end{array}$$

TABLE 4.2
Block-triangularizations for Example 4.4.

DM-decomposition of A		Combin. canon. form of $\tilde{A}$ (by p_τ)	
size	blocks	size	blocks
C_x		$C = C_w + C_x\ [R_T]$	
104	1	$114 = 75 + 39$ [75]	1
28	1	$24 = 15 + 9$ [15]	1
23	1	$18 = 10 + 8$ [10]	1
14	1	$14 = 8 + 6$ [8]	1
10	5	$6 = 4 + 2$ [4]	1
8	1	$4 = 2 + 2$ [2]	15
6	7	$2 = 1 + 1$ [1]	3
4	2		
3	9		
1	240	1	846

where ξ^i and η_i are the current in and the voltage across branch i ($i = 1, 2$). The family of minimizers of p_τ is given by

$$L(p_\tau) = \{\varnothing, \{\eta_1\}, \{\eta_2\}, \{\eta_1, \eta_2\}, \{\xi^1, \eta_1, \eta_2\}, \{\xi^2, \eta_1, \eta_2\}, \{\xi^1, \xi^2, \eta_1, \eta_2\}\}$$

and therefore the principal partition of $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$ based on p_τ yields the partition of $C = \{\xi^1, \xi^2, \eta_1, \eta_2\}$ into 4 singletons with the partial order given by $\{\eta_i\} \prec \{\xi^j\}$ ($i, j = 1, 2$). However, it is clear by inspection that $\{\eta_1, \eta_2\}$ cannot be split. On the other hand, the method using p_γ gives the partition $C = \{\xi^1\} \cup \{\xi^2\} \cup \{\eta_1, \eta_2\}$ with the partial order $\{\eta_1, \eta_2\} \prec \{\xi^i\}$ ($i = 1, 2$).

In the following, we compare the decompositions induced by the two submodular functions p_τ of (4.6) and p_γ of (4.7) associated with a layered mixed matrix $A \in \mathscr{L}\mathscr{M}(\mathbf{F}/\mathbf{K}; m_Q, m_T, n)$ of the form (3.3). Remember that $L(p)$ is defined in (2.2) as the family of minimizers of $p: 2^C \rightarrow \mathbf{R}$ and that $L(p)$ is a distributive sublattice if p is submodular.

LEMMA 5.1.

(1) $p_\tau(X) \leqq p_\gamma(X)$ *for* $X \subset C$.

(2) $\min p_\tau = \min p_\gamma$.

(3) $L(p_\tau) \supset L(p_\gamma)$.

(4) *For* $X \in L(p_\tau)$ *there exists* $Y \in L(p_\gamma)$ *such that* $Y \subset X$.

(5) $\min L(p_\tau) = \min L(p_\gamma)$.

Proof. (1) and (2): Given in (4.8) and Lemma 4.1.

(3): Immediate from (1) and (2) above.

(4): Let $Y_0 (\subset X)$ be a minimizer of $\min \{\gamma(Y) - |Y| \,|\, Y \subset X\} = \tau(X) - |X|$. From (2), we have $\min p_\gamma = \min p_\tau = \rho(X) + \gamma(Y_0) - |Y_0| \geqq \rho(Y_0) + \gamma(Y_0) - |Y_0| = p_\gamma(Y_0)$, i.e., $Y_0 \in L(p_\gamma)$.

(5): This follows from (3) and (4) above. □

In view of the correspondence between the distributive sublattices and the partition into partially ordered blocks (§2), this lemma shows that the decomposition of the column-set C (i.e., the set of variables) by the principal partition of $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$ is finer (including the partial order) than that of the combinatorial canonical form of the present paper. In other words, the column-set of each block of the combinatorial canonical form is an aggregation of the blocks of the principal partition of $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$. It is indicated by Lemma 5.1(5), however, that the column-sets of the horizontal tail are identical in both decompositions.

In Theorem 4.5 we have seen that the decomposition of C based on p_γ provides the finest block-triangular form under the equivalence transformation of the form (4.1). By a similar argument it can be shown that the principal partition of C associated with $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$ leads to the finest block-triangularization with the property (5) (as well as (2)) of Theorem 4.4, under a wider class of transformations of the following form:

$$P_r \begin{pmatrix} S_Q & 0 \\ 0 & S_T \end{pmatrix} \begin{pmatrix} Q \\ T \end{pmatrix} P_c \tag{5.1}$$

where $S_Q \in GL(m_Q, \mathbf{K})$; $S_T \in GL(m_T, \mathbf{F})$; and P_r and P_c are permutation matrices of orders m and n, respectively.

This type of transformation, however, does not seem natural and would be different from what is intended in considering a hierarchical decomposition of a system into subsystems. Recall, for instance, the matrix A of Example 5.1. Since its column-set is decomposed into singletons by $L(p_\tau)$, it can be put into a triangular form by the transformation (5.1) with $S_T = (y^{ij})^{-1}$, which could be determined only after the parameter values y^{ij} are fixed. This simple example would demonstrate that the transformation

(4.1) is more suitable in practical situations than (5.1), and hence p_γ is more appropriate than p_τ.

Note that the transformed matrix (5.1) no longer belongs to $\mathcal{LM}(\mathbf{F}/\mathbf{K}; m_Q, m_T, n)$. This suggests that the block-triangularization by the principal partition of $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$ is more adequate when considered for a broader class of matrices. This issue will be discussed in §7.

Let Γ_A and Γ_Q be defined as (4.4) respectively for A and Q. As is well known, the DM-decomposition is induced by $L(p_{\mathrm{DM}})$, where

$$p_{\mathrm{DM}}(X) = |\Gamma_A(X)| - |X| \qquad (X \subset C). \tag{5.2}$$

Since $|\Gamma_A(X)| = |\Gamma_Q(X)| + |\Gamma_T(X)| \geqq \rho(X) + \gamma(X)$, we have

$$p_\gamma(X) \leqq p_{\mathrm{DM}}(X) \qquad (X \subset C). \tag{5.3}$$

THEOREM 5.2. *If A $(\in \mathcal{LM}(\mathbf{F}/\mathbf{K}))$ is nonsingular, then*

$$\min p_\tau = \min p_\gamma = \min p_{\mathrm{DM}} = 0$$

and

$$L(p_\tau) \supset L(p_\gamma) \supset L(p_{\mathrm{DM}}).$$

Proof. The relations between p_τ and p_γ follow from Lemma 5.1. By Theorem 4.2, the assumption is equivalent to $\min p_\gamma = 0$, which, combined with (5.3) and $p_{\mathrm{DM}}(\emptyset) = 0$, yields $\min p_{\mathrm{DM}} = 0$. The inclusion $L(p_\gamma) \supset L(p_{\mathrm{DM}})$ is then evident from (5.3). □

Example 5.2. This is continued from Examples 3.1 and 4.2. As given in [30], the principal partition of $C = \{\xi^i, \eta_i | i = 1, \cdots, 9\}$ associated with $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$ consists of 10 blocks; the block $C_8 = \{\eta_5, \xi^5, \xi^9\}$ of the combinatorial canonical form in Example 4.2 splits into two blocks $\{\eta_5\}$ and $\{\xi^5, \xi^9\}$. It should be mentioned that, as opposed to the claim of [30], the unknown variables $\{\xi^5, \xi^9\}$ cannot be determined independently of η_5 even after the variables of $C_9 = \{\xi^2, \eta_2, \xi^3, \eta_3, \xi^4, \eta_4, \xi^7, \xi^8\}$ are fixed.

Example 5.3. For a singular matrix the canonical form is not a refinement of the DM-decomposition. Consider, e.g., the matrix

$$A = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{|cccc|} \hline 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 \\ \hline 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 1 & 1 \\ \hline \end{array} \end{array}, \tag{5.4}$$

which may be thought of as a member of $\mathcal{LM}(\mathbf{F}/\mathbf{Q}; 4, 0, 4)$ $(\mathbf{F} \supset \mathbf{Q})$. The canonical form consists of tails only; $C_0 = \{1, 2, 3, 4\}$, $|R_0| = 2$, $C_\infty = \emptyset$, $|R_\infty| = 2$. On the other hand, the DM-decomposition evidently decomposes A into 2 square blocks.

Example 5.4. For the problem of Example 4.3 the decompositions based on p_γ and p_τ are compared in Table 4.1.

5.2. Decomposition for electrical networks with admittance expression. In general, an electrical network can be described by the structural equations and the constitutive equations among currents ξ^i in and voltages η_i across the branches (cf. Example 3.1). When the branch characteristics are given in terms of self- and mutual-admittances Y, the coefficient matrix A of the system of equations in (ξ, η) takes the form:

$$A = \begin{array}{c} \begin{array}{cc} \xi & \eta \end{array} \\ \begin{array}{|cc|} \hline D & 0 \\ 0 & R \\ \hdashline -I & Y \\ \hline \end{array} \end{array} \tag{5.5}$$

where D and R are the fundamental cutset and circuit matrices respectively. If the nonvanishing entries of Y are assumed to be algebraically independent over $\mathbf{Q}$, the trivial transcendental scaling of the constitutive equations brings it into the class of $\mathscr{LM}(\mathbf{R}/\mathbf{Q})$. In this extended sense, we will regard A as a member of $\mathscr{LM}(\mathbf{R}/\mathbf{Q})$ of the form (3.3) with

$$Q=\begin{pmatrix} D & 0 \\ 0 & R \end{pmatrix}, \qquad T=(-I \quad Y). \tag{5.6}$$

The column-set C of A of (5.5) is the disjoint union of two copies, say B_ξ and B_η, of the set B of branches; i.e.,

$$C=B_\xi \cup B_\eta. \tag{5.7}$$

This allows us to identify the Boolean lattice 2^C with the direct product of 2^{B_ξ} and 2^{B_η}. It may also be noted that the row-set of Y is identified with B_ξ, while its column-set is B_η.

The decomposition of C proposed in [16], [17] is as follows. Let $\mu(I)$ and $\nu(I)$ denote the rank and the nullity of the arc set $I(\subset B)$ in the underlying graph. Obviously, we have

$$\mu(B\backslash J)=\nu(J)-|J|+\mu(B). \tag{5.8}$$

The nonsingularity of A of (5.5) can be formulated [42] in terms of an independent-matching problem on the bipartite graph representing Y, where the matroid with rank function μ is attached to both B_ξ and B_η. Put

$$\mathscr{H}=\{(I,J) | I\subset B_\xi, J\subset B_\eta, I\supset \Gamma_Y(J)\} \tag{5.9}$$

where Γ_Y is defined for Y as in (4.4), and

$$p_\mu(I,J)=\mu(I)+\mu(B_\eta\backslash J)-\mu(B) \qquad (I\subset B_\xi, J\subset B_\eta). \tag{5.10}$$

Note that $(I, J)\in \mathscr{H}$ iff $(I, B_\eta\backslash J)$ is a cover of Y, and then $p_\mu(I, J) + \mu(B)$ is the rank of the cover in the independent-matching problem. The set of minimizers of $p_\mu|_{\mathscr{H}}$, the restriction of p_μ to $\mathscr{H}$, is denoted simply as $L(p_\mu)$, i.e.,

$$L(p_\mu)=\{(I,J)\in\mathscr{H} | p_\mu(I,J)=\min_{\mathscr{H}} p_\mu\}, \tag{5.11}$$

which is a sublattice of $2^{B_\xi}\times 2^{B_\eta}\simeq 2^C$ (cf. (5.7)), and hence determines a decomposition of C into partially ordered blocks. We call this the decomposition by the minimum covers of the admittance matrix.

The rest of this subsection is devoted to establishing Theorem 5.4 below, which implies that the combinatorial canonical form for A of the particular form (5.5) gives an essentially identical block-triangularization with the one provided by the decomposition by the minimum covers of the admittance matrix.

From (5.8) and (5.10) we see that

$$p_\mu(I,J)=\mu(I)+\nu(J)-|J| \qquad (I\subset B_\xi, J\subset B_\eta). \tag{5.12}$$

On the other hand, p_γ of (4.7) for A of (5.5) is written as

$$\begin{aligned} p_\gamma(I\cup J)&=\rho(I\cup J)+|I\cup\Gamma_Y(J)|-|I\cup J| \\ &=\mu(I)+\nu(J)-|J|+|\Gamma_Y(J)\backslash I| \qquad (I\subset B_\xi, J\subset B_\eta), \end{aligned} \tag{5.13}$$

since the rank ρ of $\mathbf{M}(Q)$ is equal to $\mu+\nu$. Combining (5.12) and (5.13), we obtain

$$p_\gamma(I\cup J)=p_\mu(I,J)+|\Gamma_Y(J)\backslash I| \qquad (I\subset B_\xi, J\subset B_\eta). \tag{5.14}$$

LEMMA 5.3.

$$p_Y(I \cup J) = p_\mu(I,J) \quad \textit{for}\ (I,J) \in \mathscr{H},$$

$$p_Y(I \cup J) > p_\mu(I,J) \quad \textit{for}\ (I,J) \notin \mathscr{H}.$$

Proof. From (5.14) it follows that $p_Y \geqq p_\mu$, where the equality holds iff $\Gamma_Y(J) \subset I$. □

THEOREM 5.4.

(1) $\min \{p_Y(I \cup J) | I \subset B_\xi, J \subset B_\eta\} = \min \{p_\mu(I,J) | (I,J) \in \mathscr{H}\}$.

(2) $L(p_Y) \supset L(p_\mu)$.

(3) $\{J \subset B_\eta | I \subset B_\xi, I \cup J \in L(p_Y)\} = \{J \subset B_\eta | (I,J) \in L(p_\mu)\}$.

Proof. (1): By (5.13), we have

$$\begin{aligned} \min p_Y &= \min \{\min \{\mu(I) + |\Gamma_Y(J) \backslash I| \,|\, I \subset B_\xi\} + \nu(J) - |J| \,|\, J \subset B_\eta\} \\ &= \min \{\mu(\Gamma_Y(J)) + \nu(J) - |J| \,|\, J \subset B_\eta\}, \end{aligned} \tag{5.15}$$

since $\min \{\mu(I) + |\Gamma_Y(J) \backslash I| \,|\, I \subset B_\xi\} = \min \{\mu(I) + |\Gamma_Y(J) \backslash I| \,|\, I \subset \Gamma_Y(J)\} = \mu(\Gamma_Y(J))$. This establishes (1) when combined with the rather obvious relation

$$\begin{aligned} \min_{\mathscr{H}} p_\mu &= \min \{\mu(I) + \nu(J) - |J| \,|\, I \supset \Gamma_Y(J), J \subset B_\eta\} \\ &= \min \{\mu(\Gamma_Y(J)) + \nu(J) - |J| \,|\, J \subset B_\eta\}. \end{aligned} \tag{5.16}$$

(2): Immediate from Lemma 5.3 and (1) above.

(3): From (5.15) and (5.16) it is easy to see that the families on both sides of (3) agree with the minimizers $J(\subset B_\eta)$ of $\mu(\Gamma_Y(J)) + \nu(J) - |J|$. □

Theorem 5.4(2) shows that the decomposition method of the present paper applied to (5.5) yields a finer partition of the variables $\{\xi, \eta\}$ than the decomposition by the minimum covers of the admittance matrix. However, the difference is not substantial, since, as indicated by Theorem 5.4(3), they provide the identical partition for the voltage-variables η which play the primary role in (5.5); the current-variables ξ are only secondary as they are readily obtained from η by means of the admittance matrix Y. In this way, we may say that they give essentially the same decomposition. The following exemplifies that the inclusion in Theorem 5.4(2) is proper in general.

Example 5.5. For the following matrix

$$A = \begin{array}{cccc} \xi^1 & \xi^2 & \eta_1 & \eta_2 \\ \hline 1 & & & \\ & 1 & & \\ \hdashline -1 & & y^{11} & 0 \\ & -1 & y^{21} & y^{22} \end{array}, \tag{5.17}$$

the combinatorial canonical form based on $L(p_Y)$ decomposes $\{\xi^1, \xi^2, \eta_1, \eta_2\}$ into 4 singletons with the partial order:

$$\{\eta_2\} \prec \{\eta_1\} \prec \{\xi^1\}, \qquad \{\eta_2\} \prec \{\xi^2\}.$$

The decomposition by the minimum covers of Y, on the other hand, gives the partition into two blocks as

$$\{\xi^2, \eta_2\} \prec \{\xi^1, \eta_1\}.$$

6. Block-triangularization of mixed matrices. In this section, we consider the block-triangularization of a mixed matrix $A = Q_A + T_A \in \mathscr{MM}(\mathbf{F}/\mathbf{K}; m, n)$ of (3.2) under the transformation

$$SAP_c = S(Q_A + T_A)P_c, \tag{6.1}$$

where $S \in GL(m, \mathbf{K})$, and P_c is a permutation matrix. It is derived from the combinatorial canonical form of the associated layered mixed matrix $\tilde{A} \in \mathscr{LM}(\mathbf{F}/\mathbf{K};\ m,\ m,\ m + n)$ of (3.9).

Let $C_w = \{w_1, \cdots, w_m\}$ and $C_x = \{x_1, \cdots, x_n\}$ be the row-set and the column-set of A, respectively; the column-set C of $\tilde{A}$ is then identified with $C_w \cup C_x$. Suppose that the transformation (4.1) with $S_Q \in GL(m, \mathbf{K})$, and P_T, P_r and P_c permutation matrices gives the combinatorial canonical form of $\tilde{A}$ with the partition of column-set $C = \cup\{C_j | j = 0, 1, \cdots, r, \infty\}$ and the row-set $R = \cup\{R_j | j = 0, 1, \cdots, r, \infty\}$. As in §4, $\bar{Q} = S_Q[I|Q_A]P_c$ and $\bar{T} = P_T[-I|T_A]P_c$ are block-triangularized, i.e., $\bar{Q}[R_{Qi}, C_j] = 0$ and $\bar{T}[R_{Ti}, C_j] = 0$ for $i > j$, where $R_j = R_{Qj} \cup R_{Tj}$.

Put $C_{wj} = C_w \cap C_j$ and $C_{xj} = C_x \cap C_j$, and notice that the row-set R_T of $\bar{T}$ is in one-to-one correspondence with C_w. With this correspondence in mind, we have seen in (4.44) that $R_{Tj} \supset C_{wj}$ if $R_{Tj} \neq \emptyset$.

LEMMA 6.1. *Suppose* $w_k \in R_{Tj} \backslash C_{wj} (\subset C_w)$. *Then* $\{w_k\}$, *as a subset of* C, *constitutes a block, say* C_i, *in the combinatorial canonical form, where* $R_{Ti} = \emptyset$, $|R_{Qi}| = 1$, *and it is an immediate successor of* C_j, *that is,* $C_j |\!\prec C_i = \{w_k\}$.

This lemma shows that $\{\hat{C}_j = R_{Tj} \cup C_{xj},\ j \in J^*\}$, where $J^* = \{0, \infty\} \cup \{j | 1 \leqq j \leqq r,\ R_{Tj} \cup C_{xj} \neq \emptyset\}$, gives a partition of C, which is coarser than $\{C_j | j = 0, 1, \cdots, r, \infty\}$ and on which the partial order is induced from that on $\{C_j\}$ by the natural order homomorphism. Let us denote by $\{\hat{R}_j | j \in J^*\}$ the corresponding partition of R; i.e., $\hat{R}_j = R_{Qj} \cup \{R_{Qi} | C_i \subset R_{Tj} \backslash C_{wj}\} \cup R_{Tj}$. Then, by the construction, we have $|\hat{R}_{Tj}| = |\hat{C}_{wj}|$ and

$$\bar{T}[\hat{R}_{Tj}, \hat{C}_{wj}] = -I \tag{6.2}$$

where $\hat{R}_{Tj} = R_{Tj}$ and $\hat{C}_{wj} = C_w \cap \hat{C}_j$.

Since $\{\hat{C}_j\}$ and $\{\hat{R}_j\}$ are aggregations of $\{C_j\}$ and $\{R_j\}$, respectively, we have $\bar{Q}[\hat{R}_{Qi}, \hat{C}_j] = 0$ and $\bar{T}[\hat{R}_{Ti}, \hat{C}_j] = 0$ for $i > j$. If we choose $S_T = \bar{Q}[R_Q, C_w]$, we see that the matrix $\bar{Q} + S_T\bar{T}$ is block-triangularized with respect to $\{\hat{C}_j\}$ and $\{\hat{R}_j\}$, and that its submatrix corresponding to column-set C_w is the zero matrix. Denote by $\hat{A}$ the submatrix of $\bar{Q} + S_T\bar{T}$ corresponding to the column-set C_x. In view of the identity:

$$\begin{pmatrix} I_m & S_T \\ 0 & I_m \end{pmatrix} \begin{pmatrix} S_Q & \\ & P_T \end{pmatrix} \begin{pmatrix} I_m & Q_A \\ -I_m & T_A \end{pmatrix} = \begin{pmatrix} S_Q - S_TP_T & S_QQ_A + S_TP_TT_A \\ -P_T & P_TT_A \end{pmatrix}, \tag{6.3}$$

this means that the block-triangular matrix $\hat{A}$ is obtained from A by the admissible transformation of the form (6.1), since we have $S_Q = S_TP_T$, and $\hat{A} = (S_QQ_A + S_TP_TT_A)\hat{P}_c = S_Q(Q_A + T_A)\hat{P}_c = S_QA\hat{P}_c$, where $\hat{P}_c$ is a permutation matrix.

Thus we have obtained a block-triangular form of a mixed matrix A under the transformation of the form (6.1). Note that the partition of the column-set C_x of A is induced from that of the combinatorial canonical form of the corresponding layered mixed matrix $\tilde{A}$. It is easy to see from Theorem 4.5 that this is the finest block-triangularization under the transformation (6.1). Note, however, that the obtained matrix no longer belongs to $\mathscr{MM}(\mathbf{F}/\mathbf{K};\ m,\ n)$ in general. See [24] for more details.

It has been shown in [23] that if $A \in \mathscr{MM}(\mathbf{F}/\mathbf{K};\ n,\ n)$ satisfies $\det A \in \mathbf{K}\backslash\{0\}$, then there exist permutation matrices P_r and P_c, a lower triangular matrix $L \in GL(n, \mathbf{K})$ and an upper triangular matrix U over $\mathbf{F}$ such that $P_rAP_c = LU$. This result can be derived easily from the present construction if one notices (4.21).

Example 6.1. Consider the mixed matrix $A = Q_A + T_A \in \mathscr{MM}(\mathbf{F}/\mathbf{Q};\ 5,\ 5)$ given by

$$
(6.4)\qquad A = \begin{array}{r|ccccc|}
 & x_1 & x_2 & x_3 & x_4 & x_5 \\
\hline
w_1: & 1 & 1 & t_1 & 1 & t_2 \\
w_2: & -1 & -1 & 1 & t_3 & 0 \\
w_3: & 0 & 0 & t_4 & t_5 & t_6 \\
w_4: & 0 & 0 & 0 & 0 & 1 \\
w_5: & t_7 & t_8 & 0 & 0 & 0 \\
\hline
\end{array}
$$

where $\{t_i | i = 1, \cdots, 8\}$ are indeterminates over $\mathbf{Q}$, and $\mathbf{F} = \mathbf{Q}(t_1, \cdots, t_8)$. By the combinatorial canonical form of the associated layered mixed matrix $\tilde{A} \in \mathcal{LM}(\mathbf{F}/\mathbf{Q}; 5, 5, 10)$ of (3.9), we see that

$$
(6.5)\qquad P_r \begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & Q_A \\ -I & T_A \end{bmatrix} P_c
= \begin{array}{r|cccccccccc|}
 & x_1 & x_2 & w_5 & w_1 & w_2 & x_3 & x_4 & w_3 & x_5 & w_4 \\
\hline
 & 1 & 1 & & 1 & & & 1 & & & \\
w_5: & t_7 & t_8 & -1 & & & & & & & \\
 & & & 1 & & & & & & & \\
 & & & & 1 & 1 & 1 & 1 & & & \\
w_1: & & & & -1 & 0 & t_1 & 0 & & t_2 & \\
w_2: & & & & 0 & -1 & 0 & t_3 & & & \\
w_3: & & & & 0 & 0 & t_4 & t_5 & -1 & t_6 & \\
 & & & & & & & & 1 & & \\
 & & & & & & & & & 1 & 1 \\
w_4: & & & & & & & & & & -1 \\
\hline
\end{array}
$$

where

$$
(6.6)\qquad S = \begin{array}{|ccccc|}
\hline
0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
\hline
\end{array}\,.
$$

The column-set C of $\tilde{A}$, identified with $\{w_1, \cdots, w_5\} \cup \{x_1, \cdots, x_5\}$, is divided into six nonempty blocks: $C_1 = \{x_1, x_2\}$, $C_2 = \{w_5\}$, $C_3 = \{w_1, w_2, x_3, x_4\}$, $C_4 = \{w_3\}$, $C_5 = \{x_5\}$, $C_6 = \{w_4\}$ ($C_0 = C_\infty = \varnothing$) with the partial order:

$$C_1 \prec C_2;\quad C_3 \prec C_4;\quad C_1 \prec C_3 \prec C_5 \prec C_6.$$

The aggregated partition $\{\hat{C}_j | j \in J^*\}$ is given by $J^* = \{0, \infty\} \cup \{1, 3, 5, 6\}$, $\hat{C}_1 = C_1 \cup C_2 = \{x_1, x_2, w_5\}$, $\hat{C}_3 = C_3 \cup C_4 = \{x_3, x_4, w_1, w_2, w_3\}$, $\hat{C}_5 = C_5 = \{x_5\}$, $\hat{C}_6 = C_6 = \{w_4\}$ (and $\hat{C}_0 = \hat{C}_\infty = \varnothing$).

Then the following block-triangular form is obtained, where $P_c = I$:

$$
(6.7)\qquad SAP_c = \begin{array}{|ccccc|}
x_1 & x_2 & x_3 & x_4 & x_5 \\
\hline
t_7 & t_8 & & & \\
1 & 1 & t_1 & 1 & t_2 \\
 & & t_1+1 & t_3+1 & t_2 \\
 & & t_4 & t_5 & t_6 \\
 & & & & 1 \\
\hline
\end{array}\,.
$$

7. Extensions and remarks. It has been mentioned in §5.1 that the principal partition of $\mathbf{M}(Q)^* \wedge \mathbf{M}(T)$, which corresponds to the transformation (5.1), should be considered in a wider class of matrices than $\mathcal{LM}(\mathbf{F}/\mathbf{K})$. Let $\mathbf{F}_0$ be an intermediate field of $\mathbf{F}/\mathbf{K}$, $\mathbf{K} \subset \mathbf{F}_0 \subset \mathbf{F}$, and consider a matrix $A \in \mathcal{M}(\mathbf{F}; m, n)$:

$$A = \left(\begin{array}{c} Q \\ \hline T \end{array} \right), \tag{7.1}$$

such that (i) $Q \in \mathcal{M}(\mathbf{K}; m_Q, n)$, (ii) $T = Q_1 T_1 \in \mathcal{M}(\mathbf{F}; m_T, n)$ where $Q_1 \in \mathcal{M}(\mathbf{F}_0; m_T, n)$ and T_1 is a diagonal matrix of order n with its diagonal entries being algebraically independent numbers in $\mathbf{F}$ over $\mathbf{F}_0$. The class of such matrices A will be denoted by $\mathcal{LC}(\mathbf{F}/\mathbf{F}_0/\mathbf{K}; m_Q, m_T, n)$. It should be noted that $A \in \mathcal{LM}(\mathbf{F}/\mathbf{K}; m_Q, m_T, n)$ belongs to $\mathcal{LC}(\mathbf{F}/\mathbf{F}_0/\mathbf{K}; m_Q, m_T, n)$ for some $\mathbf{F}_0$, but not conversely.

It is known that the identity given in Theorem 3.1 still holds for $A \in \mathcal{LC}(\mathbf{F}/\mathbf{F}_0/\mathbf{K})$ with ρ and τ being the rank functions of $\mathcal{M}(Q)$ and $\mathcal{M}(T)$ for the submatrices in (7.1). Therefore, the partition of the column-set C based on $L(p_\tau)$, followed by appropriate row transformations, brings about a block-triangular form with the properties (1) to (5) of Theorem 4.4. Note that the block-triangular form is obtained from A by means of the transformation (5.1), where we may assume without loss of generality that $S_T \in GL(m_T, \mathbf{F}_0)$, and hence the transformed matrix remains in $\mathcal{LC}(\mathbf{F}/\mathbf{F}_0/\mathbf{K})$.

The considerations above naturally suggest an extension to multi-layered matrices of the form

$$A = \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_k \end{bmatrix} \tag{7.2}$$

such that

$$A_0 \in \mathcal{M}(\mathbf{K}; m_0, n),$$
$$A_i = Q_i T_i \in \mathcal{M}(\mathbf{F}_i; m_i, n) \qquad (i = 1, \cdots, k),$$

where

$$\mathbf{K} \subset \mathbf{F}_0 \subset \cdots \subset \mathbf{F}_k \tag{7.3}$$

is a sequence of field extensions, $Q_i \in \mathcal{M}(\mathbf{F}_{i-1}; m_i, n)$, and $T_i \in \mathcal{M}(\mathbf{F}_i; n, n)$ is a diagonal matrix with its diagonal entries being algebraically independent over $\mathbf{F}_{i-1}$ $(i = 1, \cdots, k)$. Then, by Theorem 3.1, the rank of A is expressed in terms of the rank functions ρ_i of the associated matroids $\mathbf{M}(A_i)$ $(i = 0, 1, \cdots, k)$ as

$$\operatorname{rank} A = \min \{p(X) | X \subset C\} + n \tag{7.4}$$

where

$$p(X) = \rho_0(X) + \rho_1(X) + \cdots + \rho_k(X) - |X|. \tag{7.5}$$

Based on $L(p)$, we can obtain a block-triangular canonical form with the properties (1) to (5) of Theorem 4.4 under the transformation

$$P_r \begin{bmatrix} S_0 & & & \\ & S_1 & & \\ & & \ddots & \\ & & & S_k \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_k \end{bmatrix} P_c \tag{7.6}$$

where $S_0 \in GL(m_0, \mathbf{K})$; $S_i \in GL(m_i, \mathbf{F}_{i-1})$ $(i = 1, \cdots, k)$; and P_r and P_c are permutation matrices.

The canonical form for multi-layered matrix introduced above seems to have a natural meaning for electrical networks involving multi-ports, which have been investigated in [37], [38], [39]. To be specific, consider an electrical network consisting of k multi-ports, each of which is described by a set of equations with coefficient matrix A_i $(i = 1, \cdots, k)$. Let A_0 denote the matrix (over $\mathbf{Q}$) for Kirchhoff's laws. Then the coefficient matrix for the whole system is written as (7.2) (cf. Example 3.1), and the permissible transformation (7.6) reflects the locality in the sense that we can choose an appropriate description for each device. Furthermore, the assumption of the algebraic independence among different devices would be fairly realistic.

Without the hierarchy of fields (7.3), we may likewise consider the block-triangularization based on p of (7.5) for a matrix of (7.2). That is, we may define a canonical form for a matrix A of (7.2) with $A_i \in \mathcal{M}(\mathbf{F}; m_i, n)$ $(i = 0, 1, \cdots, k)$ under the transformation (7.6) with $S_i \in GL(m_i, \mathbf{F})$ $(i = 0, 1, \cdots, k)$. In this case, however, the diagonal blocks are no longer guaranteed to be nonsingular. Two special cases may be worth mentioning. The one is the case where $k = 1$ and $A_0 = A_1$. Then the transformation (7.6), in which we may assume $S_0 = S_1$, yields the combinatorial canonical form of a matrix with respect to its pivotal transforms introduced in [14]. The other is where A is nonsingular. Then it has empty tails and the square blocks must necessarily be nonsingular.

The combinatorial canonical form introduced in this paper should prove to be a useful tool in the structural analysis of systems. For example, it is reported in [26] that it plays a central role in deriving a necessary and sufficient combinatorial condition for the structural controllability of a dynamical system described in the so-called "descriptor form": $F d\mathbf{x}/dt = A\mathbf{x} + B\mathbf{u}$, where the entries of F, A and B are assumed to be classified into accurate and inaccurate numbers in the sense of [28].

Finally, we mention the possibility of parametrizing the function p_γ as

$$p_\gamma(X; \alpha, \beta) = \alpha p(X) + \beta\gamma(X) - |X|, \qquad X \subset C.$$

According to the general framework [19], we then obtain a finer partition of the column-set of a layered mixed matrix. The significance of such a decomposition is yet to be made clear.

Acknowledgments. The authors thank Professor Satoru Fujishige of the University of Tsukuba for discussions and comments. Examples 4.3 and 4.4 are based on the data provided by Junkichi Tsunekawa of the Institute of the Japanese Union of Scientists and Engineers and on the FORTRAN program originally coded in [12] by Masaki Ichikawa of the Environment Agency, as well as on the kind help of Professor Yoshio Oyanagi of the University of Tsukuba.

REFERENCES

[1] M. AIGNER, *Combinatorial Theory*, Springer-Verlag, New York, 1979.

[2] G. BIRKHOFF, *Lattice Theory*, 3rd ed., Amer. Math. Soc. Colloq. Publ. 25, Providence, RI, 1967.

[3] W. H. CUNNINGHAM, *Matroid partition and intersection algorithms*, Dept. Math. Stat., Carleton Univ., 1984.

[4] A. L. DULMAGE AND N. S. MENDELSOHN, *Coverings of bipartite graphs*, Canad. J. Math., 10 (1958), pp. 517–534.

[5] ———, *A structure theory of bipartite graphs of finite exterior dimension*, Trans. Roy. Soc. Canada, Section III, 53 (1959), pp. 1–13.

[6] ———, *On the inversion of sparse matrices*, Math. Comp., 16 (1962), pp. 494–496.

[7] ———, *Two algorithms for bipartite graphs*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 183–194.

[8] J. EDMONDS, *Minimum partition of a matroid into independent subsets*, J. Nat. Bur. Stand., 69B (1965), pp. 67–72.

[9] ———, *Systems of distinct representatives and linear algebra*, J. Res. Nat. Bur. Stand., 71B (1967), pp. 241-245.

[10] S. FUJISHIGE, *Algorithms for solving the independent-flow problems*, J. Oper. Res. Soc. Japan, 21 (1978), pp. 189-203.

[11] F. HARARY, *A graph-theoretic approach to matrix inversion by partitioning*, Numer. Math., 4 (1962), pp. 128–135.

[12] M. ICHIKAWA, *An application of matroid theory to systems analysis* (in Japanese), Graduation thesis, Dept. Math. Engrg. Instr. Phys., University of Tokyo, 1983.

[13] M. IRI, *The maximum-rank minimum-term rank theorem for the pivotal transforms of a matrix*, Linear Algebra Appl., 2 (1969), pp. 427–446.

[14] ———, *Combinatorial canonical form of a matrix with applications to the principal partition of a graph* (in Japanese), Trans. Inst. Electr. Comm. Engrg. Japan, 54A (1971), pp. 30–37. (English translation in Electronics and Communications in Japan, 54A (1971), pp. 30–37.)

[15] ———, *A review of recent work in Japan on principal partitions of matroids and their applications*, Ann. New York Acad. Sci., 319 (1979), pp. 306–319.

[16] ———, *Application of matroid theory to engineering systems problems*, in Proc. Sixth Conf. Prob. Theory, B. Bereanu et al., eds., Editura Academiei Republicii Romania, 1981, pp. 107–127.

[17] ———, *Applications of matroid theory*, in Mathematical Programming—The State of the Art, A. Bachem, M. Grötschel and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 158–201.

[18] ———, *Application of matroid theory to systems analysis and control*, (presented at RUTCOR's Inaugural Conf., Rutgers Univ., November 1983), Res. Memo. RMI 83-06, Dept. Math. Engrg. Instr. Phys., University of Tokyo, 1983.

[19] ———, *Structural theory for the combinatorial systems characterized by submodular functions*, in Progress in Combinatorial Optimization, W. R. Pulleyblank, ed., Academic Press, New York, 1984, pp. 197–219.

[20] M. IRI AND S. FUJISHIGE, *Use of matroid theory in operations research, circuits and systems theory*, Internat. J. Systems Sci., 12 (1981), pp. 27–54.

[21] M. IRI, J. TSUNEKAWA AND K. MUROTA, *Graph-theoretic approach to large-scale systems—Structural solvability and block-triangularization* (in Japanese), Trans. Infor. Process. Soc. Japan, 23 (1982), pp. 88–95. (English translation available: Res. Memo. RMI 81-05, Dept. Math. Engrg. Instr. Phy., University of Tokyo, 1981.)

[22] K. MUROTA, *Menger-decomposition of a graph and its application to the structural analysis of a large-scale system of equations*, Kokyuroku, Res. Inst. Math. Sci., Kyoto University, 453 (1982), pp. 127–173; Discrete Appl. Math., to appear.

[23] ———, *LU-decomposition of a matrix with entries of different kinds*, Linear Algebra Appl., 49 (1983), pp. 275–283.

[24] ———, *Structural solvability and controllability of systems*, Doctor's dissertation, University of Tokyo, 1983. (Improved and augmented version to appear as a monograph in Algorithms and Combinatorics from Springer.)

[25] ———, *Combinatorial canonical form of layered mixed matrices and block-triangularization of large-scale systems of linear/nonlinear equations*, DPS 257, Inst. Socio-Economic Planning, University of Tsukuba, 1985.

[26] ———, *Refined study on structural controllability of descriptor systems by means of matroids*, SIAM J. Control Optim., 25 (1987), to appear.

[27] K. MUROTA AND M. IRI, *Matroid-theoretic approach to the structural solvability of a system of equations* (in Japanese), Trans. Infor. Process. Soc. Japan, 24 (1983), pp. 157–164.

[28] ———, *Structural solvability of systems of equations—A mathematical formulation for distinguishing accurate and inaccurate numbers in structural analysis of systems*, Japan J. Appl. Math., 2 (1985), pp. 247–271.

[29] M. NAKAMURA, *Boolean sublattices connected with minimization problems on matroids*, Math. Programming, 22 (1982), pp. 117–120.

[30] ———, *Mathematical analysis of discrete systems and its applications* (in Japanese), Doctor's dissertation, University of Tokyo, 1982.

[31] ———, *Analysis of discrete systems and its applications* (in Japanese), Trans. Inst. Electr. Comm. Engrg. Japan, J66A (1983), pp. 368-373. (English abstract in ibid. 66E (1983), p. 262.)

[32] M. NAKAMURA AND M. IRI, *Fine structures of matroid intersections and their applications*, Proc. Internat. Symp. Circuit and Systems, Tokyo, 1979, pp. 996–999.

[33] ———, *A structural theory for submodular functions, polymatroids and polymatroid intersections*, Res. Memo. RMI 81-06, Dept. Math. Engrg. Instr. Phys., University of Tokyo, 1981.

[34] O. Ore, *Graphs and matching theorems*, Duke Math. J., 22 (1955), pp. 625–639.

[35] T. Ozawa, *Common trees and partition of two-graphs* (in Japanese), Trans. Inst. Electr. Comm. Engrg. Japan, 57A (1974), pp. 383–390.

[36] ———, *Topological conditions for the solvability of linear active networks*, Internat. J. Circuit Theory Appl., 4 (1976), pp. 125–136.

[37] B. Petersen, *Investigating solvability and complexity of linear active networks by means of matroids*, IEEE Trans. Circuits and Systems, CAS-26 (1979), pp. 330–342.

[38] A. Recski, *Unique solvability and order of complexity of linear networks containing memoryless n-ports*, Internat. J. Circuit Theory Appl., 7 (1979), pp. 31–42.

[39] A. Recski and M. Iri, *Network theory and transversal matroids*, Discrete Appl. Math., 2 (1980), pp. 311–326.

[40] N. Tomizawa, *Strongly irreducible matroids and principal partition of a matroid into strongly irreducible minors* (in Japanese), Trans. Inst. Electr. Comm. Engrg. Japan, J59A (1976), pp. 83–91.

[41] N. Tomizawa and S. Fujishige, *Historical survey of extensions of the concept of principal partition and their unifying generalization to hypermatroids*, Syst. Sci. Res. Rep., No. 5, Dept. Syst. Sci., Tokyo Inst. Tech., 1982.

[42] N. Tomizawa and M. Iri, *An algorithm for determining the rank of a triple matrix product* AXB *with application to the problem of discerning the unique solution in a network* (in Japanese), Trans. Inst. Electr. Comm. Engrg. Japan, 57A (1974), pp. 834–841. (English translation in Electron. Comm. Japan, 57A (1974), pp. 50–57.)

[43] B. L. van der Waerden, *Algebra*, Springer-Verlag, Berlin, 1955.

[44] D. J. A. Welsh, *Matroid Theory*, Academic Press, London, 1976.

[45] K. Yajima, J. Tsunekawa and S. Kobayashi, *On equation-based dynamic simulation*, Proc. World Congr. Chem. Eng., Montreal, V (1981).

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

BOUNDS ON THRESHOLD DIMENSION AND DISJOINT THRESHOLD COVERINGS*

PAUL ERDÖS†, EDWARD T. ORDMAN†‡ AND YECHEZKEL ZALCSTEIN†

Abstract. The threshold dimension (threshold covering number) of a graph G is the least number of threshold graphs needed to edgecover the graph G. If tc (n) is the greatest threshold dimension of any graph of n vertices, we show that for some constant A,

$$n - A\sqrt{n}\log n < \text{tc}(n) < n - \sqrt{n} + 1.$$

We establish the same bounds for edge-disjoint coverings of graphs by threshold graphs (threshold partitions). We give an example to show there exist planar graphs on n vertices with a smallest covering of An threshold graphs and a smallest partition of Bn threshold graphs, with $B = 1.5A$. Thus the difference between these two covering numbers can grow linearly in the number of vertices.

Key words. threshold graph, threshold dimension, threshold partition, graph partition

AMS(MOS) subject classifications. 05C, 68E

1. Preliminaries. By a graph $G = (V, E)$ we mean a finite set V of vertices and a collection E of edges: distinct unordered pairs of distinct vertices. A subgraph of a graph G is a subset V' of V together with a subset E' of E that consists only of edges between vertices of V'. An *induced subgraph* of a graph is a subset of the vertices together with all edges of the original graph that connect those vertices. For further notation see [6].

If x is a vertex of a graph G, the *star* of x is the subgraph consisting of x, the edges containing x, and the other vertices contained in those edges. A *stable set* of vertices (also called an independent set) is a set of vertices which induces no edges. A *dominating set* of vertices is one such that every vertex in the graph is connected to at least one of them by an edge. If a single vertex is a dominating set, it is called a *dominating vertex*. To build a *cone* on G means to add a new vertex to V and connect it to all other vertices by edges.

Threshold graphs were introduced in [2], [3], [8]. A graph is a *threshold graph* if it meets one of the following equivalent conditions:

a) It does not have as an induced subgraph a square (C_4), two disconnected edges ($2K_2$) or a path of three consecutive edges (P_4).

b) The vertices can be labelled with integers $l(v)$, and there is an integer constant t (the threshold) such that a set $\{v_1, v_2, \cdots, v_k\}$ of vertices is stable if and only if $l(v_1) + \cdots + l(v_k) < t$.

c) The vertices can be labelled with integers $l(v)$, and there is an integer constant t (these numbers may be different than those in (b)) such that any two vertices x and y are connected by an edge if and only if $l(x) + l(y) \geqq t$.

d) Every induced subgraph of G, including G itself, has at most one nontrivial component (there may be isolated vertices) and this component has a dominating vertex.

Since every edge of G is, taken by itself, a threshold graph, every graph G may be covered by threshold graphs. The smallest number of threshold subgraphs (not necessarily induced subgraphs) of G that cover G is called the *threshold dimension* of G; we will also call it the *threshold covering* number of G and denote it by tc (G). From an applied perspective, tc (G) is the smallest number of semaphores needed to synchronize a system

* Received by the editors October 29, 1984; accepted for publication (in revised form) April 23, 1986.

† Memphis State University, Memphis, Tennessee 38152.

‡ The work of this author was partially supported by National Science Foundation grant DCR-8503922.

of parallel processes definable by the graph G using PV-chunk synchronizing primitives [8]; alternatively, it is the smallest number of 0-1 simultaneous linear inequalities which can replace such a system of linear inequalities represented by G; see [3], [7], or [6, Chap. 10]. For other prior results on tc (G), see [3].

Two subgraphs of G are called *edge-disjoint* (or simply disjoint) if they have no edges in common. Since the covering of a graph G by its edges is a covering by disjoint threshold graphs, it follows that for every graph there is defined a unique integer tp (G), the *disjoint threshold dimension* or *threshold partition* number of G, the smallest number of edge-disjoint threshold graphs that will cover G.

Since every threshold partition is a threshold covering, tp $(G) \geqq$ tc (G). One goal of this paper is to begin exploring the questions, when is tp $(G) =$ tc (G)? How different can they be? For example, for some corresponding results for clique coverings and clique partitions, see [1].

It should be noted that while it is easy to determine if G is a threshold graph (that is, if tc $(G) = 1$), determining tc (G) is in general NP-complete [3]; in fact, it is NP-complete to test if tc $(G) = 3$ [10] or even if tc $(G) = 2$ [4].

LEMMA 1. *If G is a triangle-free graph,* tc $(G) =$ tp (G).

Proof. As observed in [2], if G contains no triangle, every threshold graph contained in G is a star. Suppose G is covered by k stars $S_1, S_2, \cdots, S_k$. Define $S'_1 = S_1$, $S'_2 = S_2 - S_1$, and in general $S'_j = S_j - (S_1 \cup \cdots \cup S_{j-1})$ for $j = 2$ to k. Clearly the various S'_j are disjoint stars and cover G, so tp $(G) \leqq$ tc (G) as required.

2. The size of a required threshold covering. In [3], Chvâtal and Hammer raise the issue: how big need tc (G) be? They prove [3, Thm. 3] that if $\alpha(G)$ is the size of the largest stable set in a graph G with n vertices, then tc $(G) \leqq n - \alpha(G)$ with equality holding if G is triangle-free (and in some other cases). They also observe [3, Cor. 3A] that for every positive ε, there is a graph G on n vertices with tc $(G) > (1 - \varepsilon)n$. In fact, the proof of their Corollary 3A shows more than this. We restate it as follows:

THEOREM 1. *There is a constant A such that for large enough n there is a graph G with n vertices and*

$$\text{tp}\,(G) = \text{tc}\,(G) > n - A\sqrt{n}\log(n).$$

Proof. In [5], Erdös shows that for a sufficiently large fixed constant A, there is an integer N such that for $n > N$ there is a graph G on n vertices with no triangle and with no stable set of $A\sqrt{n}\log(n)$ vertices. Thus tp $(G) =$ tc (G), and

$$\alpha(G) < A\sqrt{n}\log(n) \quad \text{and} \quad \text{tc}\,(G) > n - A\sqrt{n}\log(n)$$

as desired.

This shows that there are graphs with relatively large values of tc (G). We now turn to improving the upper bound on tp (G).

THEOREM 2. *Let G be an arbitrary graph on n vertices. Then*

$$\text{tp}\,(G) < n - \sqrt{n} + 1.$$

Proof. Suppose there is a stable set A in G of size $\sqrt{n}$ or larger. Then Theorem 3 of [3] points out that the stars on $V - A$ provide a covering of G by no more than $n - \sqrt{n}$ threshold graphs; Lemma 1 above shows how to make this a threshold partition.

Now by contrast suppose that no stable set in G has as many as $\sqrt{n}$ elements. Pick a vertex z in G; let $x_1, \cdots, x_k$ be a maximal stable set in the star of z; hence $k < \sqrt{n}$. For each x_i, in turn, we construct a graph T_i consisting of all edges starting at x_i together with any triangles including the edge (z, x_i); omit from this any edges included in a previous T_j to keep the T_i's disjoint. (To see that T_i is threshold, use definition (c). Label

x_i with 4; z with 3; any vertex which neighbors z and x_i but no previous $x_j, j < i$, with 2; other points adjoining x_i with 1. Let $t = 5$.)

We have now constructed k edge-disjoint threshold graphs which cover the union of the stars of the $k + 1$ vertices $z, x_1, \cdots, x_k$. Delete the covered edges from G. This eliminates at least $k + 1$ vertices. Since it deletes an edge only when deleting at least one vertex on it, the reduced graph G' cannot have a bigger independent set than G had.

Reduce G' by choosing a new z. At each stage, we eliminate $k + 1$ vertices by covering them with k threshold graphs;

$$k < \sqrt{n} \quad \text{so} \quad \frac{k}{k+1} < \frac{\sqrt{n}}{\sqrt{n}+1}$$

and the total number of graphs needed to cover all n vertices is not greater than

$$\frac{n\sqrt{n}}{\sqrt{n}+1} < n - \sqrt{n} + 1$$

which completes the proof of Theorem 2.

We now let tc (n) denote the largest tc (G) for any G with n vertices; tp (G) is defined similarly. The above results show that

$$n - A\sqrt{n}\log(n) < \mathrm{tc}\,(n) < n - \sqrt{n} + 1$$

and

$$n - A\sqrt{n}\log(n) < \mathrm{tp}\,(n) < n - \sqrt{n} + 1.$$

It remains of interest to tighten these bounds, and to know whether the limits for tc (n) and tp (n) are actually the same. A private communication from János Pach [9] improves the upper bound in each case to $n - \sqrt{n \log n}$ for triangle-free graphs only.

3. The difference between tc (G) and tp (G). Since the bounds we have established for tc (G) and tp (G) are identical, it is reasonable to ask whether tc (G) and tp (G) are ever very different. Our object in this section is to show that tp (G)–tc (G) can grow proportionally to the number of vertices n in G, even if G is a planar connected graph or a very highly-connected graph of low diameter.

We will make heavy use of a threshold graph H constructed as follows: consider six vertices $x_1, \cdots, x_6$ and connect x_i and x_j if $i + j \leqq 7$. Note that the deletion of the single edge x_2x_3 would make it cease to be threshold since then $x_5x_2x_4x_3$ would be an induced path.

Example 1. Let G_{10} be the graph made by taking two copies of H and identifying the two copies of x_2, x_3, and the edge between them. This graph is shown in Fig. 1; it is planar. Clearly tc $(G_{10}) = 2$, since it is covered by two copies of H. The reader may verify that tp $(G_{10}) = 3$; two graphs in the partition are a copy of H and a path $x_4x_3x_1$. The proof that there is no partition into two threshold graphs hinges on the fact that x_2x_3 would have to be in the same graph as one "wing" x_1x_6; the side of G_{10} lacking x_2x_3 cannot then be covered by one threshold graph.

The reader may also wish to verify that G_{10} is a critical example; deleting an x_1x_6 from G_{10} results in tc = tp = 2, deleting any other edge yields tc = tp = 3.

The graph G_{10} may be used to build various examples in which the difference between tc (G) and tp (G) grows linearly in the number of vertices or edges of G. For example, if G^r is the disjoint union of r copies of G_{10}, tp $(G^r) = 3r$ and tc $(G^r) = 2r$. This example may be made planar and connected by joining successive copies G_{10} together at the "wingtips" (identify an x_6 of one G_{10} with an x_6 from another). To build more highly

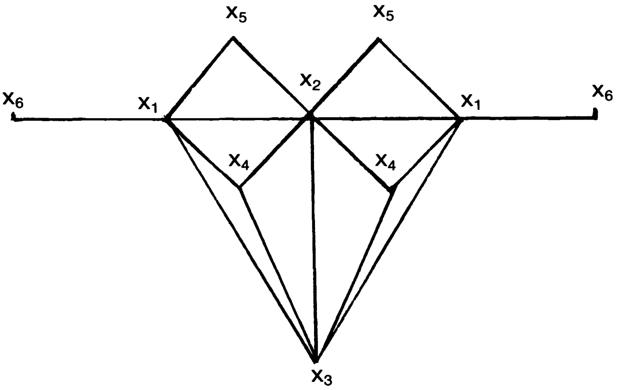


FIG. 1. *The graph* G_{10}.

connected (but nonplanar) examples, we use the following lemma motivated by a discussion with V. Chvátal:

LEMMA 2. *Let G' denote the cone on the (arbitrary) graph G. Then*

$$\operatorname{tc}(G')=\operatorname{tc}(G) \quad \textit{and} \quad \operatorname{tp}(G')=\operatorname{tp}(G).$$

Proof. Any threshold covering of G' induces a (no larger) threshold covering of G since an induced subgraph of a threshold graph is a threshold graph. Given a (disjoint) threshold cover of G, we obtain a (disjoint) threshold cover of G' by picking any threshold graph D in the cover of G and enlarging it to include the new vertex of G' and its star in G'. That the enlarged D remains a threshold graph is easily seen by definition (d) of threshold graphs; the new vertex of G' is a dominating vertex in the enlarged version of D.

Using this lemma, we can create an arbitrarily highly connected graph with tc $= 2r$, tp $= 3r$, by taking G^r and erecting a cone on it as many times as desired (that is, add 5 new points all connected to all original points and each other, to make it 5-connected).

It is now clear that there is a constant c_1 such that a graph G on n vertices can have tp (G) − tc $(G) \geqq c_1 n$. How big can c_1 be? Example G_{10} shows it can be at least $\frac{1}{10}$. What upper bound can be put on tp (G) − tc (G)? We know it cannot exceed $n - \sqrt{n} - 1$, but we believe this can be improved. Finally, can tp (G)/tc (G) ever exceed $\frac{3}{2}$? If so, how big can it be?

REFERENCES

[1] L. CACCETTA, P. ERDŐS, E. ORDMAN AND N. PULLMAN, *The difference between the clique numbers of a graph*, Ars Combin., 19A (1985), pp. 97–106.

[2] V. CHVÁTAL AND P. HAMMER, *Set packing and threshold graphs*, Univ. Waterloo Research Report CORR, 73-21 (1973).

[3] ———, *Aggregation of inequalities in integer programming*, Ann. Discrete Math., 1 (1977), pp. 145–162.

[4] M. COZZENS AND R. LEIBOWITZ, handwritten preprint, 1984.

[5] P. ERDŐS, *Graph theory and probability* II, Canad. J. Math., 13 (1961), pp. 346–352.

[6] M. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[7] P. L. HAMMER, T. IBARAKI AND U. N. PELED, *Threshold numbers and threshold completion*, in Studies on Graphs and Discrete Programming, P. Hansen, ed., North-Holland Publishing Company, Amsterdam, 1981, pp. 125–145.

[8] P. HENDERSON AND Y. ZALCSTEIN, *A graph theoretic characterization of the* PV *chunk class of synchronizing primitives*, SIAM J. Comput., 6 (1977), pp. 88–108.

[9] J. PACH, private communication, letter of October 13, 1984.

[10] M. YANNAKAKIS, *The complexity of the partial order dimension problem*, this Journal, 3 (1982), pp. 351–358.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

CHANGE OF BASIS FOR PRODUCTS OF ORTHOGONAL POLYNOMIALS*

STEPHEN BARNETT†

Abstract. Given two sets of orthogonal polynomials $\mathscr{B} = \{p_i(\lambda)\}$ and $\{q_i(\lambda)\}$, simple procedures are given for expressing the products $\lambda^i p_j(\lambda)$, $\lambda^i q_j(\lambda)$, $p_i(\lambda)q_j(\lambda)$ and $q_i(\lambda)q_j(\lambda)$ in terms of the basis $\mathscr{B}$. The main computations involved are multiplications of vectors by a tridiagonal matrix. The results are based on a previous theorem for determining the product of two polynomials both expressed relative to $\mathscr{B}$.

Key words. orthogonal polynomials, matrix methods

AMS(MOS) subject classifications. Primary 42C05; secondary 15A99

1. Introduction.

Consider a set $\mathscr{B} = \{p_i(\lambda)\}$ of orthogonal polynomials defined by the usual formulae

$$p_0(\lambda) = 1, \qquad p_1(\lambda) = \alpha_1\lambda + \beta_1, \tag{1.1}$$

$$p_i(\lambda) = (\alpha_i\lambda + \beta_i)p_{i-1}(\lambda) - \gamma_i p_{i-2}(\lambda), \qquad i = 2, 3, \cdots, \tag{1.2}$$

with $\alpha_i > 0$, and let

$$a(\lambda) = p_n(\lambda) + a_1 p_{n-1}(\lambda) + \cdots + a_n p_0(\lambda) \tag{1.3}$$

be an nth degree generalized polynomial expressed relative to the basis $\mathscr{B}$. It has been shown in a previous paper [2] that if

$$b(\lambda) = p_m(\lambda) + b_1 p_{m-1}(\lambda) + \cdots + b_m p_0(\lambda), \qquad m \leqq n \tag{1.4}$$

is a second generalized polynomial, then the product $a(\lambda)b(\lambda)$ can be determined relative to $\mathscr{B}$ by carrying out some very simple operations involving the $N \times N$ tridiagonal matrix

$$A = \begin{bmatrix} \dfrac{-\beta_1}{\alpha_1} & \dfrac{1}{\alpha_1} & 0 & 0 & & & \\ \dfrac{\gamma_2}{\alpha_2} & \dfrac{-\beta_2}{\alpha_2} & \dfrac{1}{\alpha_2} & 0 & & 0 & \\ 0 & \dfrac{\gamma_3}{\alpha_3} & \dfrac{-\beta_3}{\alpha_3} & \dfrac{1}{\alpha_3} & \ddots & & \\ & & \ddots & \ddots & \ddots & & \\ & 0 & & \ddots & & \dfrac{-\beta_{N-1}}{\alpha_{N-1}} & \dfrac{1}{\alpha_{N-1}} \\ & & & & & \dfrac{\gamma_N}{\alpha_N} & \dfrac{-\beta_N}{\alpha_N} \end{bmatrix}. \tag{1.5}$$

It is convenient here to record this result in the following form:

LEMMA. *The product of* (1.3) *and* (1.4) *is*

$$a(\lambda)b(\lambda) = \frac{\alpha_1\alpha_2\cdots\alpha_m}{\alpha_{n+1}\alpha_{n+2}\cdots\alpha_{n+m}} p_{n+m}(\lambda) + \sum_{i=0}^{m+n-1} u_i p_i(\lambda) \tag{1.6}$$

* Received by the editors November 20, 1985; accepted for publication (in revised form) April 23, 1986.

† School of Mathematical Sciences, University of Bradford, Bradford, West Yorkshire, BD7 1DP, England.

where

$$[u_0, u_1, \cdots, u_{m+n-1}] = R_{m+1} + b_1 R_m + \cdots + b_m R_1 \tag{1.7}$$

and R_i is the ith row of the matrix $a(A)$, in which A has order $N = m + n$. Furthermore, these rows are given by

$$R_1 = [a_n, a_{n-1}, \cdots, a_1, 1, 0, \cdots, 0] \tag{1.8}$$

and

$$R_i = R_{i-1}(\alpha_{i-1}A + \beta_{i-1}I) - \gamma_{i-1}R_{i-2}, \qquad i = 2, 3, \cdots \tag{1.9}$$

where I denotes the unit matrix of order N, and $\gamma_1 = 0$.

In this paper it is shown how the products $\lambda^i p_j(\lambda)$, $\lambda^i q_j(\lambda)$, $p_i(\lambda)q_j(\lambda)$ and $q_i(\lambda)q_j(\lambda)$, where $\{q_i(\lambda)\}$ is a second set of orthogonal polynomials, can all be expressed in terms of $\mathscr{B}$. The results (Theorems 1 to 5) are all derived from the lemma, and so involve only simple vector-matrix multiplications of the type occurring in (1.9). Throughout, no conversions of polynomials to power form are required. Some numerical examples emphasize the simplicity of the approach and also illustrate how sequences of products of increasing degrees are obtained.

The methods presented in this paper are more straightforward than those of Salzer [9]–[11], and form part of a continuing programme on the algebraic manipulation of generalized polynomials using matrix techniques [1]–[7].

2. The product $\lambda^i p_j(\lambda)$. In the lemma set $a(\lambda) = \lambda^i$ and $b(\lambda) = p_j(\lambda)$ to obtain:

THEOREM 1. *Let the $(j + 1)$th row of A^i be*

$$\rho_{j+1,i} = [v_0, v_1, v_2, \cdots, v_{N-1}]. \tag{2.1}$$

Then the product $\lambda^i p_j(\lambda)$ is given by

$$\lambda^i p_j(\lambda) = \sum_{k=0}^{i+j} v_k p_k(\lambda), \qquad i + j \leqq N \tag{2.2}$$

where $v_{i+j} = 1/\alpha_{j+1}\alpha_{j+2} \cdots \alpha_N$ if $i + j = N$.

Notice that when $i + j = N$ the leading coefficient in (2.2) has to be modified from that in (1.6) with $m = j$, $n = i$ since $a(\lambda) = \lambda^i$ is not monic relative to $\mathscr{B}$. Note also that the case $i = 1$ is trivial, since the expression for $\lambda p_j(\lambda)$ can be obtained directly by rearranging (1.2) in the form

$$\lambda p_j(\lambda) = [p_{j+1}(\lambda) - \beta_{j+1}p_j(\lambda) + \gamma_{j+1}p_{j-1}(\lambda)]/\alpha_{j+1}. \tag{2.3}$$

The rows in (2.1) can be computed iteratively by a formula of the type (1.9), which here becomes

$$\rho_{j+1,i} = \rho_{ji}(\alpha_j A + \beta_j I) - \gamma_j \rho_{j-1,i}, \qquad j \geqq 1 \tag{2.4}$$

and the first row of A^i can be conveniently determined from the following scheme. Write

$$\lambda^i = \sum_{k=0}^{i} t_{ik} p_{i-k}(\lambda) \tag{2.5}$$

in which $t_{i0} = 1/\alpha_1\alpha_2 \cdots \alpha_i$. Then

$$\lambda^{i+1} = \sum_{k=0}^{i} t_{ik} \lambda p_{i-k}(\lambda) \tag{2.6}$$

and substituting (2.3) into (2.6) and equating coefficients of terms in $p_i(\lambda)$ gives

$$(2.7) \qquad [t_{i+1,i+1}, t_{i+1,i}, \cdots, t_{i+1,1}] = [t_{ii}, t_{i,i-1}, \cdots, t_{i0}]A_{i+1}$$

where A_i denotes the $i \times i$ leading principal submatrix of A. Since from (1.1)

$$\lambda = \left(\frac{1}{\alpha_1}\right)p_1(\lambda) - \left(\frac{\beta_1}{\alpha_1}\right)p_0(\lambda),$$

repeated use of (2.7) enables the coefficients t_{ik} in (2.5) to be determined for $i = 2, 3, \cdots$. By Theorem 1 with $j = 0$, we then have

$$(2.8) \qquad \rho_{1i} = [t_{ii}, t_{i,i-1}, \cdots, t_{i0}, 0, \cdots, 0].$$

Example 1. Throughout the illustrative examples in this paper the basis $\mathscr{B}$ will consist of the Legendre polynomials $P_i(\lambda)$, for which

$$(2.9) \qquad \alpha_i = \frac{2i-1}{i}, \quad \beta_i = 0, \quad \gamma_i = \frac{i-1}{i}, \quad i \geqq 1.$$

Let $N = 5$, so that from (1.5) and (1.9) we can write

$$(2.10) \qquad A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{5} & 0 & \frac{3}{5} & 0 \\ 0 & 0 & \frac{3}{7} & 0 & \frac{4}{7} \\ 0 & 0 & 0 & \frac{4}{9} & 0 \end{bmatrix}.$$

The submatrices A_2, A_3, A_4 are indicated within (2.10) by dashed lines. Since $\lambda = P_1(\lambda)$, (2.7) with $i = 1$ gives

$$[t_{22}, t_{21}] = [0, 1]A_2 = [\tfrac{1}{3}, 0]$$

so that

$$\lambda^2 = \frac{1}{3}P_0(\lambda) + \frac{1}{\alpha_1\alpha_2}P_2(\lambda) = \frac{1}{3}P_0(\lambda) + \frac{2}{3}P_2(\lambda)$$

and from (2.8)

$$\rho_{12} = [\tfrac{1}{3}, 0, \tfrac{2}{3}, 0, 0].$$

Applying (2.4) with $i = 2$ produces

$$\rho_{22} = \rho_{12}A = [0, \tfrac{3}{5}, 0, \tfrac{2}{5}, 0],$$

$$\rho_{32} = \rho_{22}(\tfrac{3}{2}A) - \tfrac{1}{2}\rho_{12} = [\tfrac{2}{15}, 0, \tfrac{11}{21}, 0, \tfrac{12}{35}],$$

$$\rho_{42} = \rho_{32}(\tfrac{5}{3}A) - \tfrac{2}{3}\rho_{22} = [0, \tfrac{6}{35}, 0, \tfrac{23}{45}, 0],$$

so that from Theorem 1 with $i = 2, j \leqq 3$ we obtain

$$\lambda^2 P_1(\lambda) = \frac{3}{5}P_1(\lambda) + \frac{2}{5}P_3(\lambda),$$

$$\lambda^2 P_2(\lambda) = \frac{2}{15}P_0(\lambda) + \frac{11}{21}P_2(\lambda) + \frac{12}{35}P_4(\lambda),$$

$$\lambda^2 P_3(\lambda) = \frac{6}{35} P_1(\lambda) + \frac{23}{45} P_3(\lambda) + \frac{1}{\alpha_4 \alpha_5} P_5(\lambda),$$

and $1/\alpha_4\alpha_5 = \frac{20}{63}$.

For the next step, return to (2.7) with $i = 2$ to obtain

$$[t_{33}, t_{32}, t_{31}] = [\tfrac{1}{3}, 0, \tfrac{2}{3}] A_3 = [0, \tfrac{3}{5}, 0]$$

so that

$$\lambda^3 = \frac{3}{5} P_1(\lambda) + \frac{1}{\alpha_1 \alpha_2 \alpha_3} P_3(\lambda) = \frac{3}{5} P_1(\lambda) + \frac{2}{5} P_3(\lambda)$$

and hence

$$\rho_{13} = [0, \tfrac{3}{5}, 0, \tfrac{2}{5}, 0].$$

Applying (2.4) with $i = 3$ produces ρ_{23} and ρ_{33}, leading to the expressions for $\lambda^3 P_1(\lambda)$ and $\lambda^3 P_2(\lambda)$; and so on.

3. Change of basis. Suppose that a second set of orthogonal polynomials $q_0(\lambda)$, $q_1(\lambda)$, $q_2(\lambda)$, $\cdots$ is defined by

$$q_0(\lambda) = 1, \qquad q_1(\lambda) = \delta_1 \lambda + \epsilon_1, \tag{3.1}$$

$$q_j(\lambda) = (\delta_j \lambda + \epsilon_j) q_{j-1}(\lambda) - \phi_j q_{j-2}(\lambda), \qquad j \geqq 2 \tag{3.2}$$

with $\phi_1 = 0$ and $\delta_j > 0$. We wish to express a given generalized polynomial

$$f(\lambda) = \sum_{i=0}^{\nu} f_i q_i(\lambda)$$

in terms of the basis $\mathscr{B}$. This has applications to the evaluation of integrals [8], [12]. It is sufficient to consider the case $f(\lambda) = q_j(\lambda)$:

THEOREM 2. *The first row e_{1j} of $q_j(A)$ is generated by*

$$e_{1j} = e_{1,j-1}(\delta_j A + \epsilon_j I) - \phi_j e_{1,j-2}, \qquad j \geqq 2 \tag{3.3}$$

with $e_{10} = e_1$, the first row of I. Moreover, if

$$e_{1j} = [w_0, w_1, \cdots, w_{N-1}]$$

then

$$q_j(\lambda) = \sum_{k=0}^{j} w_k p_k(\lambda), \qquad j \leqq N \tag{3.4}$$

where $w_j = \delta_1 \delta_2 \cdots \delta_N / \alpha_1 \alpha_2 \cdots \alpha_N$ if $j = N$.

Proof. Replace λ by A in (3.2), multiply the resulting identity on the left by e_1, and use the fact that $q_{j-1}(A)$ commutes with A, to obtain

$$e_1 q_j(A) = e_1 q_{j-1}(A)(\delta_j A + \epsilon_j I) - \phi_j e_1 q_{j-2}(A)$$

which is the desired expression (3.3), since $e_{1j} = e_1 q_j(A)$. The formula (3.4) then follows by setting $a(\lambda) = q_j(\lambda)$ and $b(\lambda) = 1$ in the lemma. Again, a modification is necessary when $j = N$ since $q_N(\lambda)$ is not monic with respect to $\mathscr{B}$. □

Notice that e_{11} is the first row of $\delta_1 A + \epsilon_1 I$, namely

$$[\epsilon_1 - \delta_1 \beta_1 / \alpha_1, \delta_1 / \alpha_1, 0, 0, \cdots, 0] \tag{3.5}$$

which corresponds to the obvious expression

$$q_1(\lambda) = (\epsilon_1 - \delta_1 \beta_1 / \alpha_1) p_0(\lambda) + (\delta_1 / \alpha_1) p_1(\lambda).$$

Example 2. Let the set $\{q_i(\lambda)\}$ be the Hermite polynomials $H_i(\lambda)$ for which $H_1(\lambda) = 2\lambda$, $\delta_i = 2$, $\epsilon_i = 0$, $\phi_i = 2(i-1)$. From (3.5) the first row of $H_1(A)$ is

$$e_{11} = [0, 2, 0, 0, 0].$$

From (3.3) the first row of $H_2(A)$ is, using A in (2.10),

$$e_{12} = e_{11}(2A) - 2e_{10} = [-\tfrac{2}{3}, 0, \tfrac{8}{3}, 0, 0] \tag{3.6}$$

whence by (3.4) $H_2(\lambda) = -\frac{2}{3}P_0(\lambda) + \frac{8}{3}P_2(\lambda)$. Continuing this process gives

$$e_{13} = e_{12}(2A) - 4e_{11} = [0, -\tfrac{36}{5}, 0, \tfrac{16}{5}, 0],$$

$$H_3(\lambda) = -\tfrac{36}{5}P_1(\lambda) + \tfrac{16}{5}P_3(\lambda),$$

$$e_{14} = e_{13}(2A) - 6e_{12} = [-\tfrac{4}{5}, 0, -\tfrac{160}{7}, 0, \tfrac{128}{35}],$$

$$H_4(\lambda) = -\tfrac{4}{5}P_0(\lambda) - \tfrac{160}{7}P_2(\lambda) + \tfrac{128}{35}P_4(\lambda);$$

and finally

$$e_{15} = e_{14}(2A) - 8e_{13} = [0, \tfrac{264}{7}, 0, -\tfrac{448}{9}, 0],$$

$$H_5(\lambda) = \tfrac{264}{7}P_1(\lambda) - \tfrac{448}{9}P_3(\lambda) + \tfrac{256}{63}P_5(\lambda)$$

since for $H_5(\lambda)$, $w_5 = \delta_1\delta_2\delta_3\delta_4\delta_5/\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5 = \frac{256}{63}$, where the α_i are defined in Example 1.

In fact, if we write

$$q_j(\lambda) = \sum_{k=0}^{j} q_{jk}p_k(\lambda) \tag{3.7}$$

then it is easy to obtain a recurrence formula for the coefficients q_{jk}. Comparing (3.4) and (3.7) shows that

$$e_{1j} = q_{j0}e_1 + q_{j1}e_2 + \cdots + q_{jj}e_{j+1} \tag{3.8}$$

where e_j denotes the jth row of I. When (3.8) is substituted into (3.3), we need the fact, seen by inspection of (1.5), that

$$e_kA = \frac{1}{\alpha_k}(\gamma_k e_{k-1} - \beta_k e_k + e_{k+1}), \qquad k \geqq 1. \tag{3.9}$$

Now replace e_{1j}, $e_{1,j-1}$ and $e_{1,j-2}$ in (3.3) by the expressions obtained from (3.8). On using (3.9) and equating coefficients of e_{k+1} we obtain

$$q_{jk} = \frac{\delta_j}{\alpha_k}q_{j-1,k-1} + \frac{\delta_j\gamma_{k+2}}{\alpha_{k+2}}q_{j-1,k+1} - \phi_j q_{j-2,k} + \left(\epsilon_j - \frac{\delta_j\beta_{k+1}}{\alpha_{k+1}}\right)q_{j-1,k} \tag{3.10}$$

and this five-term recurrence formula is identical to one given in [12]. However, our rederivation of (3.10) is interesting since it implies that Theorem 2 can be regarded as a convenient form of (3.10) for computational purposes.

4. Change of basis for products. The original lemma shows how the product $p_i(\lambda)p_j(\lambda)$ can be expressed in terms of $\mathscr{B}$. We now extend this to the products $\lambda^i q_j(\lambda)$, $p_i(\lambda)q_j(\lambda)$ and $q_i(\lambda)q_j(\lambda)$.

THEOREM 3. *If* e_{1j} *is as defined in Theorem* 2, *and*

$$e_{1j}A^i = [x_0, x_1, \cdots, x_{N-1}]$$

then

$$\lambda^i q_j(\lambda) = \sum_{k=0}^{i+j} x_k p_k(\lambda), \qquad i+j \leqq N \tag{4.1}$$

where $x_{i+j} = \delta_1\delta_2 \cdots \delta_j/\alpha_1\alpha_2 \cdots \alpha_N$ *if* $i + j = N$.

Proof. In the lemma set $a(\lambda) = \lambda^i q_j(\lambda)$ and $b(\lambda) = 1$ to obtain (4.1) from (1.6), again taking care that when $i + j = N$ the leading coefficient in the sum is appropriately modified. □

THEOREM 4. *The ith row of* e_{ij} *of* $q_j(A)$ *satisfies the relation*

$$e_{ij} = e_{i-1,j}(\alpha_{i-1}A + \beta_{i-1}I) - \gamma_{i-1}e_{i-2,j}, \qquad i \geqq 2 \tag{4.2}$$

with e_{1j} *defined in Theorem* 2. *Moreover, if*

$$e_{i+1,j} = [y_0, y_1, \cdots, y_{N-1}]$$

then

$$p_i(\lambda)q_j(\lambda) = \sum_{k=0}^{i+j} y_k p_k, \qquad i+j \leqq N \tag{4.3}$$

where $y_{i+j} = \delta_1\delta_2 \cdots \delta_j/\alpha_{i+1}\alpha_{i+2} \cdots \alpha_N$ *if* $i + j = N$.

Proof. Since $q_j(A)$ is a polynomial in A, the recurrence formula (4.2) follows immediately from (1.9). Setting $a(\lambda) = q_j(\lambda)$ and $b(\lambda) = p_i(\lambda)$ in the lemma reduces (1.6) to (4.3), with an appropriately modified leading coefficient when $i + j = N$. □

THEOREM 5. *Let* e'_{1i} *denote the row vector consisting of the first* $i + 1$ *elements of* e_{1i}, *and let* B *denote the* $(i + 1) \times N$ *matrix consisting of the first* $i + 1$ *rows of* $q_j(A)$. *If*

$$e'_{1i}B = [z_0, z_1, \cdots, z_{N-1}]$$

then

$$q_i(\lambda)q_j(\lambda) = \sum_{k=0}^{i+j} z_k p_k(\lambda), \qquad i+j \leqq N; \quad i \leqq j \tag{4.4}$$

where $z_{i+j} = (\delta_1\delta_2 \cdots \delta_i)(\delta_1\delta_2 \cdots \delta_j)/\alpha_1\alpha_2 \cdots \alpha_N$ *if* $i + j = N$.

Proof. Setting $a(\lambda) = q_i(\lambda)q_j(\lambda)$ and $b(\lambda) = 1$ in the lemma shows that the desired expression (4.4) for $q_i(\lambda)q_j(\lambda)$ is obtained from the elements in the first row of the matrix $q_i(A)q_j(A)$, in other words from $e_{1i}q_j(A)$. However, since $q_i(\lambda)$ has degree i, Theorem 2 implies that only the first $i + 1$ elements of e_{1i} are nonzero, so that product $e_{1i}q_j(A)$ can be replaced by that in the statement of the theorem, where B consists of the rows e_{1j}, $e_{2j}, \cdots, e_{i+1,j}$. □

Notice that to construct e_{1i} in Theorem 5 required $(i - 1)$ applications of the recurrence formula (3.3), and to construct the rows of B requires $(j - 1)$ applications of (3.3), followed by i applications of (4.2). The recurrence formulae thus need to be used a total of $2i + (j - 2)$ times, which explains why in general in Theorem 5 it will be preferable to take $i \leqq j$.

The results in Theorems 4 and 5 are particularly appealing because of the rather nice way in which the recurrence formulae for the two sets of orthogonal polynomials are intertwined via (3.3) and (4.2). As with all the procedures presented in this paper, the main computational effort arises only from the multiplication of row vectors by the tridiagonal matrix A, and the algorithms are simpler than those in [11].

Example 3. Continue with the Hermite and Legendre polynomials of Example 2. Using e_{12} in (3.6) and A in (2.10) we can readily compute

$$e_{12}A^2 = [\tfrac{2}{15}, 0, \tfrac{20}{21}, 0, \tfrac{32}{35}]$$

so by (4.1)

$$\lambda^2 H_2(\lambda) = \tfrac{2}{15}P_0(\lambda) + \tfrac{20}{21}P_2(\lambda) + \tfrac{32}{35}P_4(\lambda)$$

and similarly

$$e_{12}A^3 = [0, \tfrac{18}{35}, 0, \tfrac{44}{45}, 0]$$

giving

$$\lambda^3 H_2(\lambda) = \tfrac{18}{35}P_1(\lambda) + \tfrac{44}{45}P_3(\lambda) + \tfrac{32}{63}P_5(\lambda)$$

since in Theorem 3 $x_5 = \delta_1\delta_2/\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5 = \frac{32}{63}$, where the α_i and δ_i are defined in Examples 1 and 2, respectively.

Next, from (4.2) and (3.6) we obtain the second row of $H_2(A)$

$$e_{22} = e_{12}A = [0, \tfrac{2}{5}, 0, \tfrac{8}{5}, 0]$$

so from (4.3)

$$P_1(\lambda)H_2(\lambda) = \tfrac{2}{5}P_1(\lambda) + \tfrac{8}{5}P_3(\lambda).$$

Returning to (4.2) with $i = 3, 4$ and $j = 2$ we obtain

$$e_{32} = \tfrac{3}{2}e_{22}A - \tfrac{1}{2}e_{12} = [\tfrac{8}{15}, 0, \tfrac{2}{21}, 0, \tfrac{48}{35}],$$

$$e_{42} = \tfrac{5}{3}e_{32}A - \tfrac{2}{3}e_{22} = [0, \tfrac{24}{35}, 0, \tfrac{2}{45}, 0]$$

so by Theorem 4 we obtain, respectively,

$$P_2(\lambda)H_2(\lambda) = \tfrac{8}{15}P_0(\lambda) + \tfrac{2}{21}P_2(\lambda) + \tfrac{48}{35}P_4(\lambda),$$

$$P_3(\lambda)H_2(\lambda) = \tfrac{24}{35}P_1(\lambda) + \tfrac{2}{45}P_3(\lambda) + \tfrac{80}{63}P_5(\lambda)$$

where the last term comes from $\delta_1\delta_2/\alpha_4\alpha_5 = \frac{80}{63}$.

The procedure can be continued, using (4.2) with $j = 3$ and e_{13} in Example 2, to obtain e_{23} and e_{33} and hence the expressions for $P_1(\lambda)H_3(\lambda)$ and $P_2(\lambda)H_3(\lambda)$.

Finally, to illustrate Theorem 5 we consider the product $H_2(\lambda)H_3(\lambda)$. We need to evaluate

$$e'_{12}\begin{bmatrix} e_{13} \\ e_{23} \\ e_{33} \end{bmatrix} \tag{4.5}$$

where from (3.6) $e'_{12} = [-\frac{2}{3}, 0, \frac{8}{3}]$, e_{13} is given in Example 2 and e_{23} and e_{33} are determined as just described above. The product (4.5) is then found to be $[0, -\frac{24}{35}, 0, -\frac{512}{45}, 0]$, so from (4.4)

$$H_2(\lambda)H_3(\lambda) = -\tfrac{24}{35}P_1(\lambda) - \tfrac{512}{45}P_3(\lambda) + \tfrac{256}{63}P_5(\lambda)$$

where the last coefficient is $\delta_1\delta_2\delta_1\delta_2\delta_3/\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5$.

REFERENCES

[1] S. Barnett, *Polynomials and Linear Control Systems*, Marcel Dekker, New York, 1983.

[2] ———, *Manipulation of generalised polynomials using matrices*, Proc. International Conference on Linear Algebra and its Applications, Vitoria, Spain, 1983, pp. 9–19.

[3] ———, *Division of generalized polynomials using the comrade matrix*, Linear Algebra Appl., 60 (1984), pp. 159–175.

[4] S. BARNETT, *Multiplication of generalized polynomials, with applications to classical orthogonal polynomials*, this Journal, 5 (1984), pp. 457–462.

[5] ———, *Solution of $ax + by = d$ for generalized polynomials*, Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, 1984, pp. 1766–1767.

[6] ———, *Solution of diophantine equations for generalized polynomials*, Univ. Bradford Mathematical Sciences, Report TAM84-18, 1984.

[7] ———, *Some further results on the algebraic manipulation of generalized polynomials*, Univ. Bradford Mathematical Sciences, Report TAM85-28, 1985.

[8] R. K. LITTLEWOOD AND V. ZAKIAN, *Numerical evaluation of Fourier integrals*, J. Inst. Math. Appl., 18 (1976), pp. 331–339.

[9] H. E. SALZER, *A recurrence scheme for converting from one orthogonal expansion into another*, Comm. ACM, 16 (1973), pp. 705–707.

[10] ———, *Converting interpolation series into Chebyshev series by recurrence formulas*, Math. Comp., 30 (1976), pp. 295–302.

[11] ———, *A variable coefficient extension of a formula for series conversion*, J. Comput. Phys., 23 (1977), pp. 82–85.

[12] B. Y. TING AND Y. L. LUKE, *Conversion of polynomials between different polynomial bases*, IMA J. Numer. Anal., 1 (1981), pp. 229–234.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

003

ON FOUR PROBLEMS IN GRAPH THEORY*

ELLIS L. JOHNSON† AND SEBASTIANO MOSTERTS†

Abstract. The four problems considered are: the Chinese postman problem, the co-postman problem, the odd cut problem, and the odd circuit problem. Relationships are developed between these problems using results from dual matroids and blocking clutters. Connections with Gomory's group problem are shown. The notion of representations of these binary group problems on augmented graphs is developed along with a discussion of the class of augmented graphs having the same solution set. After some blocking and duality results, we give forbidden augmented minors for problems of one type (e.g., Chinese postman) to be also a second type of problem (e.g., odd cut). Some results are given on b-regular problems and are used in the forbidden augmented minor characterizations.

Key words. Chinese postman problem, graphs, matroids, duality

AMS(MOS) subject classifications. 05B35, 05B40, 05C50

1. Four problems. A *graph* is an undirected graph, $G = (V, E)$, which may not be connected and which may have duplicate edges, loops, and isolated nodes. That is, an edge $e \in E$ is an unordered pair of nodes $e = [i, j]$ with $i, j \in V$ but with no restrictions on what pairs of nodes constitute the edge set. Let $c: E \rightarrow R_+$ be a nonnegative cost function. We refer to $c(e)$ as the *cost of edge e*. The *cost of a set S* of edges is defined to be $\sum c(e)$ summed over $e \in S$.

Problem 1. In the *Chinese postman problem* we are given a cost function c and a given subset of the nodes $U \subset V$, called *odd nodes*. Before stating the problem, let us define the degree $d_i(S)$ of a node i for a subset S of edges to be the total number of times an edge $e \in S$ includes the node i. A loop $e = [i, i]$ includes the node i twice. Then, the Chinese postman problem is to find a minimum cost subset S of edges such that $d_i(S)$ is an odd integer for $i \in U$ and an even integer otherwise. A set S of edges satisfying the above condition on $d_i(S)$ is called a *postman set*. In order that there exist a postman set, every connected component must be *even*, that is, must contain an even number of odd nodes. We make that assumption in order to avoid having to consider infeasible problems.

The original version of this problem came from the problem of finding a minimum cost postman tour in a graph. A *tour* of a graph is a path, not necessarily simple, which returns to its origin. A *postman tour* is a tour which uses every edge at least once. The problem of finding a minimum cost postman tour, following the Mei-Ko Kwan development [10], is equivalent to the special case of the above described problem where the graph G is connected and odd nodes U are those nodes having odd degree for the entire edge set E. Then, the edges $e \in S$ in a postman set are the edges which have to be traversed twice in a postman tour. In fact, if the edges in a postman set are duplicated in the graph, then the resulting graph has an Euler tour, because it has even degree and is assumed to be connected, which is the desired postman tour of the original graph. There is a good algorithm [1] for solving this problem.

Problem 2. The *odd cut problem* has the same data given as for the Chinese postman problem: a cost function $c(e)$, $e \in E$, and a designation of a subset U of the nodes as odd nodes. Define a *cut* to be a set of edges whose removal from G would increase the number of connected components and which is minimal with respect to this property. Define an *odd cut* to be a cut which has a nonempty intersection with every postman set, for the same designated set of odd nodes. Otherwise, a cut is an *even cut*. Alternatively, we could

* Received by the editors August 28, 1985; accepted for publication (in revised form) May 9, 1986.

† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

call a cut an odd cut if its removal leads to an infeasible Chinese postman problem, that is, one with an odd connected component (an odd number of odd nodes). Then, there are necessarily two odd components. The odd cut problem is to find a minimum cost odd cut. Padberg and Rao [12] gave a good algorithm for solving this problem. Their method involves finding a minimum cut and then changing it to an odd cut.

Problem 3. For the *co-postman problem*, we are given a subset D of the edges E, called *odd edges*. The edges in $E \backslash D$ are *even edges*. The problem is to find a minimum cost subset S of edges such that in the remaining graph, with edge set $E \backslash S$, there are no odd circuits, where a *circuit* is a node-simple tour and an *odd circuit* is one containing an odd number of odd edges.

When every edge of G is considered to be odd, then the co-postman problem is to remove a minimum cost set of edges so that the remaining graph is bipartite (has no odd length circuits). This problem is equivalent to finding a maximum weight bipartite subgraph of a graph and is known to be an NP-complete problem [5].

Problem 4. The *odd circuit problem* is, simply, to find the minimum cost odd circuit in a graph, where odd circuit is defined as in Problem 3. This problem has a good algorithm [5] by contrast with the co-postman problem.

The main purpose of this paper is to establish connections between these four problem classes and to investigate their intersections. As a preliminary, the Chinese postman and co-postman problems can be restated in a more symmetric fashion.

Let us consider first the Chinese postman problem. The degree constraints as given in Problem 1 can be thought of in terms of cuts. Each node defines a cut, namely the edges meeting the node, provided the node is not a cut node. The degree constraints, then, say that the subset S of edges must meet certain odd cuts (those given by one odd node) an odd number of times and must meet certain even cuts (those given by one even node) an even number of times. The set S will then meet every odd cut an odd number of times and every even cut an even number of times. However, there is another way to define a postman solution. Take any spanning forest, that is, a spanning tree of each connected component. Then each edge f of the spanning forest can be associated with a cut consisting of the edge f and every edge in that connected component whose insertion into the tree causes a circuit containing the edge f. The edges f whose associated cut is an odd cut form a postman set, but in fact the problem can be defined as a problem of finding a set of edges which meets these odd cuts an odd number of times and these even cuts an even number of times. The edges f of the spanning forest whose associated cut is odd form a particular postman solution, and the odd cuts are precisely those cuts containing an odd number of the edges of the particular postmen solution.

In a similar way, for any spanning forest the edges out of the forest form a circuit when adjoined to the forest. Some of these circuits are odd (if they contain an odd number of odd edges) and the rest are even. The out-of-forest-edges which form odd circuits are a co-postman solution, and we obtain an equivalent co-postman problem by considering them to be the odd edges. That is, every odd circuit (using the original odd set of edges) will contain an odd number of edges of this particular co-postman set and every even circuit will contain an even number of edges of this particular co-postman set. The co-postman solutions are those sets of edges which intersect correctly (even or odd) these circuits formed by out-of-forest edges.

Thus, we see a duality in that the Chinese postman problem requires an even or odd intersection with a fundamental set of cuts whereas the co-postman problem requires edge sets having even or odd intersections with a fundamental set of circuits. However, the Chinese postman problem is better understood both from an algorithmic and polyhedral point of view, and in fact the co-postman problem is, in general, hard to solve. Somehow, intersections with cuts give an easier problem than intersections with circuits.

Mei-Ko Kwan [10] showed that one can get from one Chinese postman solution to any other by a sequence of interchanging edges in the solution and not in the solution around a circuit. In the previous terminology, we can interchange odd and even edges around a circuit and get the same problem, and any designation of odd edges giving the same solutions can be reached in this way. For co-postman problems [3], odd and even edges are interchanged on a cut.

2. Binary group problems and binary clutters. Gomory's development [4] of the group problem relates to these problems in that each of our four problems can be viewed as special cases of the group $m\mathscr{C}_2$. They were posed in that way by Gastou and Johnson [3]. We first review their development.

A *binary group* is the group $m\mathscr{C}_2$ whose elements can be represented as all 0-1 vectors of length m with addition taken modulo 2. For a subset $\mathscr{M}$ of the elements of $m\mathscr{C}_2$, let M be the 0-1 matrix with m rows and a column corresponding to each element $g \in \mathscr{M}$. The *binary group problem* is to minimize ct subject to

$$Mt \equiv b(\bmod 2), \qquad t \geqq 0 \text{ and integer,}$$

where c is a nonnegative real n-vector of costs, M is an $m \times n$ 0-1 matrix, b is a m-vector of 0's and 1's, and t is an n-vector of variables.

A binary group problem is a Chinese postman problem when M is the node-edge incidence matrix of a graph. In order to pose the other problems as binary group problems, we need to develop some duality notions.

Any binary matrix M can be brought to *standard form* $[I\,N]$ by pivot steps using modulo 2 addition; that is, by elementary row operations consisting here of adding (modulo 2) some rows to other rows. Any rows consisting of all 0's can be deleted. For an *augmented matrix* $[M \mid b]$, we bring it to standard form without pivoting on the b-column: $[I\,N \mid \tilde{b}]$. The columns in I are called *basic columns.* If any row is all 0's except in the b-column, then the corresponding binary group problem is infeasible. Thus, for any feasible binary group problem, we can bring it to this standard form. For a Chinese postman problem in standard form, the columns in the identity I correspond to edges of a spanning forest of G, the right-hand side $\tilde{b}$ tells which of those edges in the spanning forest should be equal to one in a postman solution, and the columns of N have entries of one corresponding to edges in the spanning forest in the circuit formed by adjoining an edge out of the forest to the forest.

The dual matrix to $M = [I\,N]$ is the matrix $M^* = [N^T\,I]$ of size $m^* \times n$ where $m^* = n - m$. It has the property that every row has inner product zero with every row of M.

A matrix M is *graphic* if it can be brought to the form of a node-edge incidence matrix of graph by elementary row operations. Alternatively, it is graphic if in standard form $[I\,N]$ there is some forest with edges corresponding to column of I such that the columns of N correspond to paths in the forest. There is a forbidden minor characterization of Tutte [15] for graphic matrices. A matrix is *co-graphic* if it is the dual of a graphic matrix. The co-postman problem is precisely the binary group problem with constraints

$$M^*t^* \equiv b^*$$

where $M^* = [N^T\,I]$ is cographic. The columns in I correspond to edges out of a spanning forest and the columns of N^T correspond to cut sets in the graph. We thus obtain a co-postman problem by taking the dual of a graphic matrix and any right-hand side unrelated to the odd nodes of a postman problem.

The dual of an augmented matrix $[M \mid b] = [I\, N \mid \tilde{b}]$ is the matrix

$$[M^* \mid b^*] = \left[\begin{array}{cc|c} N^T & \mathrm{I} & 0 \\ \tilde{b}^T & 0 & 1 \end{array}\right].$$

In this form, the right-hand side $b^* = (0, \cdots, 0, 1)^T$ is basic, and the binary group problem

$$M^*t^* \equiv b^*$$

is feasible if, and only if, the original b is not all 0's. The form above, with the right-hand side b^* is the basis, is called *right-hand form*, as compared to standard form. Of course, we could pivot on the bottom row to bring the right-hand side out of the basis.

For a given binary group problem with constraints

$$t \geqq 0 \text{ and integer}, \qquad Mt \equiv b \pmod 2,$$

we get another problem, called its *blocking problem*:

$$t^* \geqq 0 \text{ and integer}, \qquad M^*t^* \equiv b^* \pmod 2,$$

where $[M^* \mid b^*]$ is the dual matrix of $[M \mid b]$. The odd cut problem is the blocking problem of the Chinese postman problem, and the odd circuit problem is the blocking problem of the co-postman problem [2].

We now turn to another way of representing these problems: binary clutters. A *clutter* is simply a family of nonnested sets. Given any family of sets, we can always form a clutter from it by removing all sets which are supersets of other sets in the family. That is, the minimal sets in any family form a clutter. Given a clutter Q of subset of E, its *blocking clutter* [2] is Q^* defined by

$$Q^* = \{A^* \subseteq E \mid A^* \cap A \neq \emptyset \text{ for all } A \in Q,\ \text{and } A^* \text{ is minimal with respect to this property}\}.$$

It is clear that the blocking clutter of the blocking clutter is the original clutter.

A clutter Q is called a *binary clutter* if the cardinality of $A \cap A^*$ is odd for every $A \in Q$ and $A^* \in Q^*$, its blocking clutter. Obviously, Q is a binary clutter if and only if Q^* is. Lehman [7] gave several results on binary clutters, and Seymour [13] named them binary clutters and gave additional characterizations of them including a forbidden minor characterization. Lehman's prototype was source-to-sink paths as members of Q and source-sink separating cuts as member of Q^*. Lehman [8] refers to the members of the clutter as *ports of a matroid.*

Lehman's results [7] are in terms of binary matroids, not binary group problems. His results can be restated as follows. Given a binary group problem with augmented matrix $[M \mid b]$, form the 0-1 matrix Q of the minimal rows among all rows that are formed as row sums of M, taken modulo 2, such that the corresponding sum in the right-hand side column b is 1. This Q is a 0-1 matrix whose rows are incidence vectors of a binary clutter and every binary clutter is formed in this way. The blocking clutter is the clutter formed in this way from the blocking problem of $[M \mid b]$. These results are due to Lehman [7], but a development of them in this form can be found in [3].

We now give an example illustrating the four problems and their binary clutters. Consider the graph shown in Fig. 1. The Chinese postman problem (for odd nodes 1, 2,

$$\left[\begin{array}{ccccccc|c} 1 & 1 & 1 & & & & & 1 \\ 1 & & & 1 & & 1 & & 1 \\ & 1 & & & 1 & 1 & & 0 \\ & & & 1 & & & 1 & 1 \\ & & 1 & & 1 & & 1 & 1 \end{array}\right]$$

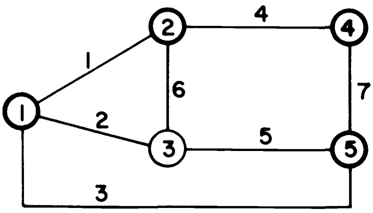


FIG. 1

4, 5) has constraints $Mt \equiv b$ where $[M \mid b]$ is the augmented matrix in Fig. 1. In standard form, it and its blocking problem (in right-hand form) are

$$\left[\begin{array}{ccccccc|c} 1 & & & & & 1 & 1 & 0 \\ & 1 & & & 1 & 1 & & 0 \\ & & 1 & & 1 & & 1 & 1 \\ & & & 1 & & & 1 & 1 \end{array}\right] \left[\begin{array}{ccccccc|c} & 1 & 1 & & 1 & & & 0 \\ 1 & 1 & & & & 1 & & 0 \\ 1 & & 1 & 1 & & & 1 & 0 \\ & & 1 & 1 & & & & 1 \end{array}\right].$$

Their associated binary clutters Q and Q^* have incidence matrices

$$Q = \begin{bmatrix} 1 & 1 & 1 & & & & \\ 1 & 1 & & 1 & 1 & & \\ 1 & & 1 & & 1 & 1 & \\ 1 & & & 1 & & 1 & \\ & 1 & 1 & & & 1 & 1 \\ & & 1 & & 1 & & 1 \\ & & & 1 & & & 1 \end{bmatrix}, \quad Q^* = \begin{bmatrix} & & 1 & 1 & & & \\ 1 & & & 1 & 1 & 1 & \\ 1 & & & & & & 1 \\ & 1 & & 1 & 1 & & \\ & 1 & & & & 1 & 1 \\ & & 1 & & 1 & 1 & 1 \end{bmatrix}.$$

The above binary clutters are blocking clutters that are the odd cuts and the *postman sets* (solutions to the Chinese postman problem). We see that the rows of the clutter matrix Q of odd cuts are solutions to the blocking problem $[M^* \mid b^*]$, i.e., the odd cut problem, and the rows of the clutter matrix Q^* of postman sets are solutions to the original problem $[M \mid b]$, i.e. the Chinese postman problem.

The co-postman problem and its blocking problem are defined by the augmented matrices $[M \mid b]$ and $[M^* \mid b^*]$

$$\left[\begin{array}{ccccccc|c} & 1 & 1 & & 1 & & & 1 \\ 1 & 1 & & & & 1 & & 1 \\ 1 & & 1 & 1 & & & 1 & 0 \end{array}\right] \left[\begin{array}{ccccccc|c} 1 & & & & & 1 & 1 & 0 \\ & 1 & & & 1 & 1 & & 0 \\ & & 1 & & 1 & & 1 & 0 \\ & & & 1 & & & 1 & 0 \\ & & & & 1 & & & 1 \end{array}\right].$$

We take a right-hand side giving as odd circuits those circuits with an odd number of edges. The blocking binary clutters Q and Q^* of odd circuits and co-postman sets are

$$\begin{bmatrix} & 1 & 1 & & 1 & & \\ 1 & 1 & & & & 1 & \\ 1 & 1 & & 1 & 1 & & 1 \\ & 1 & 1 & 1 & & 1 & 1 \end{bmatrix} \begin{bmatrix} & & & & 1 & 1 & \\ & & 1 & & & 1 & 1 \\ & 1 & & & & & \\ & & 1 & 1 & & 1 & \\ 1 & & & 1 & 1 & & \\ 1 & & & & 1 & & 1 \\ 1 & & 1 & & & & \end{bmatrix}.$$

We conclude this section with a brief review of some polyhedral results. Given the two binary group problems

$$\begin{array}{ll} t \geqq 0 \text{ and integer}, & t^* \geqq 0 \text{ and integer}, \\ Mt \equiv b \,(\text{mod } 2), & M^*t^* \equiv b^* \,(\text{mod } 2), \\ \min z = ct, & \min z^* = c^*t^*, \end{array}$$

where $[M^* \mid b^*]$ is the blocking matrix of $[M \mid b]$, we can form the two blocking clutters Q and Q^* with their corresponding incidence matrices Q and Q^* as in the previous section. Gomory's corner polyhedra [4] are given by (see [3])

$$\begin{aligned} P(M,b) &= \text{conv}\,\{t \geqq 0 \text{ and integer} \mid Mt \equiv b\,(\text{mod } 2)\} \\ &= \text{conv}\,\{q^* \mid q^* \text{ a row of } Q^*\} + R_+^n, \\ P(M^*, b^*) &= \text{conv}\,\{t^* \geqq 0 \text{ and integer} \mid M^*t^* \equiv b^*\,(\text{mod } 2)\} \\ &= \text{conv}\,\{q \mid q \text{ a row of } Q\} + R_+^n. \end{aligned}$$

Define $[M \mid b]$ to have the *Fulkerson property* [3] if

$$P(M,b) = \{t \geqq 0 \mid Qt \geqq 1\}.$$

Fulkerson [2] showed that $[M \mid b]$ has the property if and only if $[M^* \mid b^*]$ does, and he refers to $P(M, b)$ and $P(M^*, b^*)$ as blocking pairs of polyhedra. We refer to a given problem, or simply the associated augmented matrix $[M \mid b]$, as having the Fulkerson property. Fulkerson's work was based on the earlier work of Lehman [9], which concerned itself with the clutters rather than the polyhedra. Lehman [9] gave several equivalent conditions on the clutters Q and Q^* for $[M \mid b]$ to have the Fulkerson property. In general, co-postman problems and odd circuit problems do not have the Fulkerson property, but Chinese postman problems do [1] and hence so do odd cut problems [2].

3. Minors and majors. Given a binary matrix M, a *minor* $\bar{M}$ of M is another binary matrix obtained by sequentially performing two operations:

Deletion of a column of M means simply leaving it out;

Contraction of a column of M is performed by pivoting on a column and then deleting the row and column pivoted on.

In case we are trying to contract a column of all 0's, we obviously cannot pivot on the column, and in that case contraction of the column means just deleting it. On the other hand, if we delete a column which has the only nonzero in some row, then we could delete the resulting row of 0's and deletion is the same as contraction.

For our purposes and in order to be precise, let us first bring M to standard form $M = [I \mid N]$. For a matrix in standard form, the columns in I are called *basic columns* and the columns in N are called *nonbasic columns.* We get the same minors of M by restricting deletion to nonbasic columns and contraction to basic columns. To contract basic column i, we leave out the ith row and the ith column. If we want to contract a nonbasic column, we must first bring it into the basis by pivoting on any 1, which can be done unless the column is all 0's in which case it can be deleted rather than contracted but with the same effect.

For a graphic matrix M, deletion of a column gives a minor whose corresponding graph is formed by deleting the edge corresponding to the deleted column. Contracting a column of M corresponds to contracting an edge: identifying its two nodes as one node. If we contract an edge in a triangle, we cause duplicate edges to appear in the minor; if we contract an edge having a duplicate edge, then a loop appears in the minor. However, a loop has a column of all 0's in the matrix.

Our convention is to delete loops but not to contract them and to contract cut edges (one edge cut sets) but not to delete them. This convention has no effect on the matrices $\bar{M}$ which can be obtained as minors of M.

Define a *feasible minor* of an augmented matrix $[M \mid b]$ to be a minor $[\bar{M} \mid \bar{b}]$ such that

(i) b is not deleted or contracted;
(ii) The image $\bar{b}$ of b in the minor is not all 0's;
(iii) There is a 0-1 solution $\bar{t}$ to $\bar{M}\bar{t} \equiv \bar{b}$ (mod 2).

A *feasible contraction* is a feasible minor formed by contractions and no deletions, and a *feasible deletion* is a feasible minor formed by deletions only. We have two special forms of binary matrices, the first of which is standard form $[M \mid b] = [I\, N \mid b]$. In this form, we form a feasible minor by contracting a subset of the columns of I and deleting a subset of the columns of N (but not b) such that not all of the rows where $b_i = 1$ have the corresponding column I^i contracted. Condition (iii) is automatically satisfied. For a feasible deletion, both conditions (ii) and (iii) are always satisfied.

In right-hand form

$$[M^* \mid b^*] = \left[\begin{array}{cc|c} N^T & I & 0 \\ b^T & 0 & 1 \end{array}\right],$$

a feasible minor is performed by deleting nonbasic columns or by contracting columns of I. In contracting basic columns, the right-hand side should not be contracted, and in deleting nonbasic columns not all columns of b^T having $b_i^T = 1$ can be deleted in order that condition (iii) is satisfied. Condition (ii) is always satisfied in this case.

Given a feasible minor $[\bar{M} \mid \bar{b}]$ of an augmented matrix $[M \mid b]$, the blocking matrix $[\bar{M}^* \mid \bar{b}^*]$ of $[\bar{M} \mid \bar{b}]$ is a feasible minor of the blocking augmented matrix $[M^* \mid b^*]$ of $[M \mid b]$ (see [14], also [3]). This result should be clear from the previous discussion and the well-known corresponding theorem [16] for minors of dual matrixes. The latter theorem follows from the fact that deletion (contracting) of a column in M gives a minor of M whose dual is obtained by contracting (deleting) the same column of M^*, the dual of M.

Our interest is in augmented matrices $[M \mid b]$ which are either graphic or co-graphic after deleting or contracting b. The graph can be augmented as follows. For $[M \mid b]$ graphic after deleting b (the Chinese postman problem), we can bring M to the form of a node-edge incidence matrix and indicate b by designating the node i to be even if $b_i = 0$ and odd if $b_i = 1$. However, while the Chinese postman problem and its blocking odd cut problem can be treated in this way, the co-postman and odd circuit problems cannot be represented by a graph with some odd nodes and the rest even nodes. A more general procedure is to bring M to standard form $[I\, N \mid b]$, so that the edges in I, forming a spanning tree, for which $b_i = 1$ are called odd edges and indicated in figures by being drawn darker than the other edges. That is to say, take any particular Chinese postman solution and consider its edge set to be the odd edges. This way of viewing the problem was discussed at the end of § 1. A cut is even or odd depending on whether it includes an even or an odd number of edges of the particular postman solution, i.e., of odd edges. Any spanning forest including the odd edges determines a fundamental set of cuts, some of which are even and some of which are odd. A set of edges is a postman solution if, and only if, it intersects every odd cut of this fundamental set an odd number of times and every even cut of this fundamental set an even number of times. Thus, for a Chinese

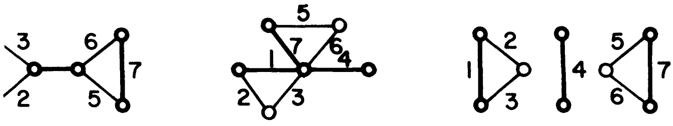


FIG. 2

postman problem and its associated blocking problem, an odd cut problem, the problem can be represented either by designating some subset of the nodes as odd nodes or by designating some subset of the edges as odd edges. The odd edge set should not include a circuit, i.e., should be a forest, and if it contained a circuit we could just remove the edges in the circuit to obtain an equivalent problem.

For co-postman and odd circuit problems, we only know the second way of specifying the problem. That is, given a graph and a subset of edges called odd edges, an odd circuit is a circuit containing an odd number of odd edges. The odd edges can now be assumed to be in the complement of some spanning forest, and each odd edge induces an odd circuit when adjoined to that spanning forest. The even edges outside of the spanning forest induce even circuits. The set of circuits, some odd and some even, formed in this way constitute a fundamental set of circuits of the graphic matroid. The co-postman sets are the sets of edges meeting the odd circuits of this fundamental set an odd number of times and the even circuits an even number of times. Thus, a co-postman problem and its blocking odd circuit problem are specified by a graph and a subset of edges outside of some spanning forest, i.e., not containing a cut set. If the set of odd edges included a cut set, then we could remove the cut set from the odd edge set and obtain an equivalent problem.

The first question we address is: which graphs and odd subset of edges give the same problem? To be specific, let us discuss Chinese postman problems. The remarks apply equally to the other three problem classes. When we say "the same problem" we mean that the clutter Q^* of solutions is equal under permutation of the rows and columns. For example, all three graphs in Fig. 2 give the same problem. This example illustrates the fact that strongly connected components can be treated separately. When we take the odd edge representation of the problem, the odd edges stay the same and the strongly connected components can be connected or disconnected in any manner provided they are not connected so as to change the strongly connected components.

Graphs that are strongly connected and not isomorphic may have the same Chinese postman solutions. It should be clear that the graphs in Figs. 3(a) and (b) are not isomorphic because one has a degree four node and the other does not. However, they do have the same Chinese postman solutions. Whitney [17] called such graphs 2-isomorphic. The construction is to take any two node cut set (the two middle nodes of Fig. 3) and "turn over" one of the two disconnected pieces. More precisely, if i and j disconnect G into H_1 and H_2, then in H_2 connect all edges that were connected to i to j instead, and vice

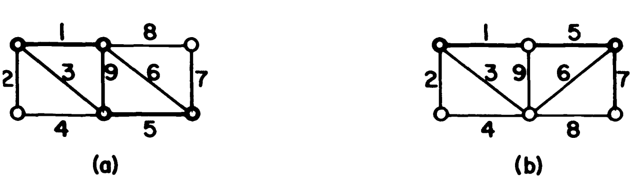


FIG. 3

versa. Define this operation to be a *2-flip on i, j.* Two graphs are 2-isomorphic if one can be reached from the other by a finite sequence of such interchanges. He showed, as is easy to see, that 2-isomorphic graphs have the same circuits. Since we have preserved one particular Chinese postman solution and since the graphs have exactly the same circuits and by Mei-Ko Kwan's result (see [10]), 2-isomorphic graphs have the same Chinese postman solutions. Whitney [17] showed the converse as well; a result that seems to be deep and not easy to prove: every two strongly connected graphs with the same sets of circuits (allowing reordering the edges) are 2-isomorphic. Thus, the class of graphs that represent a graphic matrix M is those graphs obtained from one by moving around strongly connected components and substituting any 2-isomorphic graph in place of a component.

Lehman [7, (46)] gives the result that the clutter Q determines the matrix $[M \mid b]$ provided that Q has no zero columns, which can be shown to be equivalent to the matrix $[M \mid b]$, including the right-hand side b, being *nonseparable*, i.e. $[M \mid b]$ is not block diagonal when brought to standard form. Note that M could be separable, but then the right-hand side column b must have nonzero entries for each separable component. A Chinese postman problem is nonseparable if and only if there is no strongly connected component with no odd edges. Since circuits are completely contained in a strongly connected component, interchanging odd and even edges around a circuit will always maintain some odd edges in every strongly connected component. When we "move around strongly connected components," i.e. connect or disconnect the strongly connected components in such a way that the strongly connected components remain the same, the odd and even nodes may change, but the odd and even edges remain the same. In essence, one must fix a particular Chinese postman solution before moving around the strongly connected components, and the odd and even nodes are determined so as to be consistent with that particular solution. Similar remarks apply to the co-postman problem.

THEOREM 3.1. *Two augmented graphs (graphs with odd edges) have the same sets of Chinese postman solutions (allowing reordering the edges) if and only if one can be brought to the other by sequences of the following three operations:*

(i) *Interchanging the odd and even edges around any circuit;*

(ii) *Moving around any strongly connected component;*

(iii) *Making a 2-flip on any two-node cut set.*

Proof. Let us emphasize that the odd edges do not change (but odd nodes might) under a 2-flip (see Fig. 3). The "if" direction should be clear from the previous discussion, so we only prove the other direction.

Let G and H be any two augmented graphs with the same clutter Q of Chinese postman solutions. By Lehman's result [7], they have the same sets of circuits as well. Since they have the same circuits, they are 2-isomorphic, by Whitney's theorem [17]. Since G and H have the same Chinese postman clutters Q, we can bring the odd edges of one to be the same as the other using step (i). Since G and H are 2-isomorphic, we can bring one to be the other by steps (ii) and (iii), completing the proof.

Since the blocking clutter Q^* is uniquely determined by Q, the same results hold by replacing Chinese postman by odd cut in the statement. For the odd circuit problem, there is one obvious difference: we interchange on cuts rather than circuits. Otherwise, the theorem is the same because the circuits uniquely determine the cuts. Thus, we have the theorem below.

THEOREM 3.2. *Two augmented graphs have the same sets of odd circuits (allowing reordering of the edges) if and only if one can be brought to the other by a sequence of the following three operations:*

(i) *Interchanging the odd and even edges in a cut;*

(ii) *Moving around any strongly connected component*;

(iii) *Making a 2-flip on any two-node cut set.*

By the same remarks applied to the odd cut problem, the same theorem holds for "odd circuits" replaced by "co-postman solutions."

When we say that a problem is, e.g., a co-postman problem, we are really talking about an augmented matrix $[M \mid b]$ which is co-graphic after deleting b. The graph we draw for it is the graph that M is co-graphic with respect to. An odd set of edges is the subset of basic edges of M (i.e., in a co-basis of the graph) which have a one in the right-hand side position when brought to standard form. When we form the blocking odd-circuit problem $[M^* \mid b]$, the augmented matrix is graphic after contracting b^*. The odd edges (the same as for the co-postman problem) can now be thought of as the nonbasic edges having a one in the row that is deleted after bringing b^* into the basis (or b^* may already be in the basis if the problem is in right-hand form).

That is to say, when we refer, for example, to a class of problems that is odd cut but not Chinese postman, we are not referring to the graph, but to the augmented matrix $[M \mid b]$ that, in this case, must be graphic after contracting b but not graphic after deleting b. An augmented graph only becomes meaningful when we say which problem class it represents because the same augmented graph is used for both the Chinese postman problem and its blocking odd cut problem, and could even represent an odd circuit and a co-postman problem if the odd edges do not contain a cut.

Consider now taking a feasible minor of an augmented matrix and what happens to its augmented matrix. For a Chinese postman (and odd cut) problem, the problem can be represented by an augmented graph having some nodes odd and the rest even. For this type of augmented graph, minors are formed by just deleting any edge not a cut edge, by our convention, or by contracting any edge not a loop. When contracting an edge, the new node is odd if the edge met one odd node and one even node, and the new node is even otherwise. By not allowing deletion of cut edges, the problem cannot become infeasible.

When the augmented graph is represented by having the edges in a particular solution be designated as odd edges and the other edges are even, any even edge other than a cut edge can be deleted since it can be made nonbasic. Any odd edge or even edge not a loop can be contracted. However, in this case we also allow changes of the type in Theorem 3.1(i), i.e., interchanging even and odd edges around a circuit. In fact, any of the changes in Theorem 3.1 can be made because we are really thinking of the augmented graph as representing an augmented matrix. Theorem 3.1 is stated for this representation of an augmented graph, i.e. with some odd edges forming a particular solution. When operation 3.1(ii) is done, strongly connected components can be moved around or just made separate connected components. In drawing forbidden minors we resolve the ambiguity of how to connect up the strongly connected components by drawing them as separate connected components. What we are saying is that these are Chinese postman problems on these connected components so that putting together, in any way, one solution for each component gives a solution to the Chinese postman problem. For example, in Fig. 2 the three problems are equivalent. Note that the odd node designation may change but the odd edge set does not.

DEFINITION 3.3. *An augmented minor of the graph of Chinese postman problem* is a graph obtained from a given augmented graph, with odd edges representing a particular Chinese postman solution containing no circuits by the following five operations:

(i)–(iii) As in Theorem 3.1;

(iv) Deleting any even edge that is not a bridge;

(v) Contracting any odd edge that is not the only odd edge; or

(vi) Contracting any even edge that does not form a circuit when adjoined to the odd edges.

Condition (iv) requires that the edge not be a bridge (an edge whose removal increases the number of connected components) because of our convention of not deleting edges in every basis. Similarly, condition (v) insures that we keep a nonzero right-hand side and that we contract edges in a basis. One could, of course, interchange odd edges around circuits as in (i) before doing (v).

The graph for an odd cut problem is the same as for its blocking Chinese postman problem so 3.3(i)–(v) also define augmented minors of graphs representing odd cut problems.

DEFINITION 3.4. *An augmented minor of the graph of a co-postman problem* is a graph obtained from a given augmented graph, with odd edges representing a particular co-postman solution containing no cut, by the following five operations:

(i)–(iii) As in Theorem 3.2;

(iv) Contracting any even edge not a loop;

(v) Deleting any odd edge that is not the only odd edge;

(vi) Deleting any even edge that is not a bridge in the subgraph of even edges.

4. Duality relationships between four problems. In this section we introduce the relationships between the four problems defined in the previous sections. We study their blocking connections and begin to show the dualities between them. In the next sections this subject will be discussed in more detail.

Let us consider the class of group problems with an associated augmented binary matrix $[M \mid b]$ which is graphic or co-graphic after deletion or contraction of the right-hand side b. This class can be split into fifteen regions which represent all of the possible intersections between postman, odd cut, odd circuit and co-postman problems—it is clear that for every problem belonging to any region there is a one-to-one correspondence with its blocking problem that might, or not, belong to the same region. However, all of the blocking problems of problems in any one region belong to only one region. If the blocking problems are in the same region as the original ones, we refer to this class as a self-blocking region.

THEOREM 4.1. *The self-blocking regions are those defined by the following intersections of problem classes*:

1) *Postman, odd cut, odd circuit, and co-postman*;

2) *Co-postman and odd circuit but neither postman nor odd cut*;

3) *Postman and odd cut but neither odd circuit nor co-postman.*

For example, if the augmented matrix is graphic but not co-graphic after deletion of b and co-graphic but not graphic after contraction of b then the blocking matrix $[M^* \mid b^*]$ is co-graphic but not graphic after contraction and graphic but not co-graphic after deletion.

We give here an example for each of the two first cases. The augmented matrix

$$\left[\begin{array}{cccccc|c} 1 & & & 1 & & 1 & 1 \\ & 1 & & 1 & 1 & & 1 \\ & & 1 & & 1 & 1 & 1 \end{array}\right]$$

is K_4 after deleting b so is both graphic and co-graphic, and it is $K^*_{2,3}$, the dual of $K_{2,3}$, after contracting b. The four augmented graphs associated with it are shown in Fig. 4. Thus, this augmented matrix is all four problems but on four different augmented graphs.

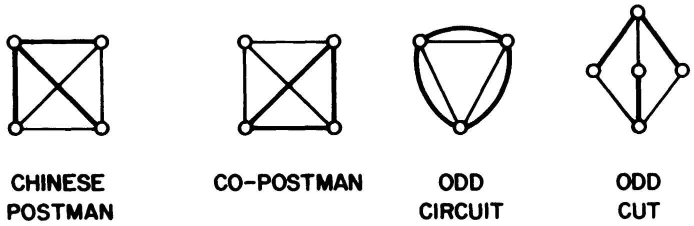


FIG. 4

A matrix giving case 2 is

$$\begin{array}{cccccccccc|c}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & \\
\hline
1 & & & & & & 1 & & 1 & & 0 \\
& 1 & & & & & 1 & & & 1 & 0 \\
& & 1 & & & & & 1 & 1 & & 0 \\
& & & 1 & & & & 1 & & 1 & 0 \\
& & & & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
& & & & & 1 & & & 1 & 1 & 1
\end{array}.$$

As an odd circuit problem, its associated augmented graph is shown in Fig. 15 and is called G_{13} there. As a co-postman problem we must pivot column six into the basis in place of the right-hand side. The associated augmented graph is K_5 with two odd edges. Since neither of the two graphs is planar, it should be clear that $[M \mid b]$ is neither a Chinese postman nor an odd cut problem. An example of Case 3 is an odd $K_{3,3}$ graph as discussed after Theorem 6.5.

In general it is well known that the class of Chinese postman problems is the blocker of the class of odd-cut problems and that the class of the co-postman problems is the blocker of the class of the odd circuit problems. Theorem 4.1 is a refinement of this fact. For example, a problem that is neither postman nor odd cut must have a blocking problem that is neither postman nor odd cut.

The problems belonging to the first region could be called b-planar problems because, whether we delete the right-hand side b or contract it, we get a matrix that is the incidence matrix of a planar graph.

We now turn our attention to planar problems after deletion or in other words to problems that are both Chinese postman and co-postman. This region can be clearly split into four subregions corresponding to odd circuit problems, or odd cut problems, or neither of them or both of them. This last region is the b-planar region already mentioned. The blocking region to the problems that are neither odd cut nor odd circuit is the odd cut and odd circuit problems. Examples of such problems are given by the Chinese postman problem on G_1^*, G_2^*, or G_3^*. Examples of the other two classes are given by

$$\left[\begin{array}{cccccccccc|c}
1 & & & & & & 1 & & 1 & & 1 \\
& 1 & & & & 1 & & 1 & & & 1 \\
& & 1 & & & 1 & & 1 & & 1 & \\
& & & 1 & & & 1 & & 1 & 1 & \\
& & & & 1 & 1 & & & 1 & & 1
\end{array}\right]
\left[\begin{array}{ccccccccc|c}
1 & & & & & & & 1 & & 1 \\
& 1 & & & & & & & 1 & 1 \\
& & 1 & & & 1 & 1 & 1 & & 1 \\
& & & 1 & & 1 & 1 & & 1 & 1 \\
& & & & 1 & 1 & & & & 1
\end{array}\right].$$

The corresponding planar graphs are shown in Figs. 5(a) and (c). They are G_{10}^* and G_6^* in Figs. 16 and 10. The black nodes represent the odd nodes associated with the corre-

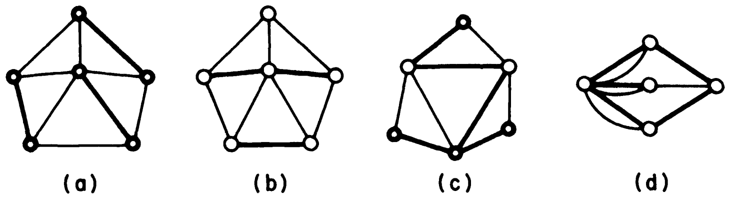


FIG. 5

sponding Chinese postman problem, and the darker edges correspond to postman sets, which are, in these two cases, spanning trees. Since the graphs are planar, the problems are also co-postman problems on the dual graphs in Figs. 5(b) and (d). Note that odd edges correspond in the primal and dual graphs. Contracting the right-hand side of the first matrix gives $K_{3,3}^*$, so it is an odd cut problem but not an odd circuit problem. Contracting the right-hand side of the second matrix gives K_5, so it is an odd circuit problem, but is not an odd cut problem.

It is easy to find examples for the remaining six regions. We give here only the blocking relationships: only odd circuit problems have blocking problems that are only co-postman problems; odd-circuit and postman problems have blocking problems that are co-postman and odd cut problems; only postman problems have blocking problems that are only odd cut problems.

5. Regular problems. A matrix M is regular if it does not contain an F_7 or F_7^* minor. Define an augmented matrix $[M \mid b]$ (and its associated binary group problem) to be *deletion regular* if M is regular and *contraction regular* if $[M \mid b]$, after contracting b, is regular. Define $[M \mid b]$ to be *b-regular* if it is both deletion regular and contraction regular. Clearly, a Chinese postman problem or a co-postman problem is deletion regular, while an odd cut problem or an odd circuit problem is contraction regular.

Consider first the case of an F_7 minor in M. It is interesting, for cases to be considered later, to ask what augmented minors of $[M \mid b]$ might be present. There are only two. The first is

$$\left[\begin{array}{cccccccc|c} 1 & & & & 1 & & 1 & 1 & 0 \\ & 1 & & & 1 & 1 & & 1 & 0 \\ & & 1 & & & 1 & 1 & 1 & 0 \\ & & & 1 & & & & & 1 \end{array}\right],$$

which we call an *even F_7 minor* of $[M \mid b]$. The fourth column cannot be contracted because it would lead to a zero right-hand side. Neither can it be deleted because a zero row with a nonzero right-hand side would result. The other possibility is that

$$\left[\begin{array}{ccccccc|c} 1 & & & 1 & & 1 & 1 & b_1 \\ & 1 & & 1 & 1 & & 1 & b_2 \\ & & 1 & & 1 & 1 & 1 & b_3 \end{array}\right]$$

is an augmented minor of $[M \mid b]$, where b_1, b_2, b_3 are not all zero. However, by pivoting and reordering rows and columns, we can bring the minor to the form M_1 below:

$$M_1 = \left[\begin{array}{ccccccc|c} 1 & & & 1 & & 1 & 1 & 1 \\ & 1 & & 1 & 1 & & 1 & 1 \\ & & 1 & & 1 & 1 & 1 & 1 \end{array}\right].$$

In fact, any right-hand side can be assumed, other than all zeros.

The case of F_7^* is similar but there are three forbidden augmented minors:

$$\text{even } F_7^* \left[\begin{array}{cccccccc|c} 1 & & & & & 1 & 1 & & 0 \\ & 1 & & & & 1 & & 1 & 0 \\ & & 1 & & & & 1 & 1 & 0 \\ & & & 1 & & 1 & 1 & 1 & 0 \\ & & & & 1 & & & & 1 \end{array}\right],$$

$$M_2 = \left[\begin{array}{ccccccc|c} 1 & & & & 1 & 1 & & 0 \\ & 1 & & & 1 & & 1 & 0 \\ & & 1 & & & 1 & 1 & 0 \\ & & & 1 & 1 & 1 & 1 & 1 \end{array}\right],$$

$$M_3 = \left[\begin{array}{ccccccc|c} 1 & & & & 1 & 1 & & 1 \\ & 1 & & & 1 & & 1 & 1 \\ & & 1 & & & 1 & 1 & 1 \\ & & & 1 & 1 & 1 & 1 & 1 \end{array}\right].$$

The minor M_2 includes the case of either one or three of the right-hand side elements b_1, b_2, b_3, b_4 equal to one. The minor M_3 includes the case of either two or all four of the b_i's equal to one. Thus we have proven the first assertion of the theorem below.

THEOREM 5.1. *The problem $[M \mid b]$ is deletion regular if, and only if, it does not include any even F_7 or even F_7^* minor or any of M_1, M_2, M_3 as augmented minors. It is contraction regular if, and only if, it does not include any even F_7 or even F_7^* minor or any of*

$$M_1^* = \left[\begin{array}{ccccccc|c} 1 & 1 & & 1 & & & & 0 \\ & 1 & 1 & & 1 & & & 0 \\ 1 & & 1 & & & 1 & & 0 \\ 1 & 1 & 1 & & & & 1 & 0 \\ 1 & 1 & 1 & & & & & 1 \end{array}\right],$$

$$M_2^* = \left[\begin{array}{ccccccc|c} 1 & 1 & & 1 & 1 & & & 0 \\ 1 & & 1 & 1 & & 1 & & 0 \\ & 1 & 1 & 1 & & & 1 & 0 \\ & & & 1 & & & & 1 \end{array}\right],$$

$$M_3^* = \left[\begin{array}{ccccccc|c} 1 & 1 & & 1 & 1 & & & 0 \\ 1 & & 1 & 1 & & 1 & & 0 \\ & 1 & 1 & 1 & & & 1 & 0 \\ 1 & 1 & 1 & 1 & & & & 1 \end{array}\right]$$

as augmented minors. It is b-regular if, and only if, it does not include any even F_7 or even F_7^* minor or any of M_1, M_2, M_3, M_1^*, M_2^*, M_3^*.

Proof. The proof is obvious from the preceding discussion and from duality.

THEOREM 5.2. *A Chinese postman problem is b-regular if, and only if, the associated graph does not contain any of the three augmented minors in Fig. 6.*

Proof. Because it is a Chinese postman problem, it is deletion regular. Hence, it cannot contain an even F_7 or even F_7^* minor or any of M_1, M_2, M_3. The graphs given are derived from M_1^*, M_2^*, and M_3^* by pivoting to standard form.

THEOREM 5.3. *An odd cut problem is b-regular if and only if its blocking Chinese postman problem is b-regular, i.e. if, and only if, its associated augmented graph does not contain G_1^*, G_2^*, or G_3^* as augmented minors.*

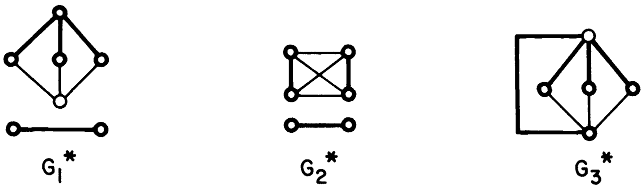


FIG. 6

For an odd cut problem $[M^* \mid b^*]$, the augmented matrix

$$M_1 = \left[\begin{array}{ccccccc|c} 1 & & & 1 & & 1 & 1 & 1 \\ & 1 & & 1 & 1 & & 1 & 1 \\ & & 1 & & 1 & 1 & 1 & 1 \end{array}\right]$$

is an example of one that can be an augmented minor of $[M^* \mid b^*]$ but is forbidden in order for $[M^* \mid b^*]$ to be a Chinese postman problem. The associated augmented graph for the odd cut problem by contracting the right-hand side of M_1 is G_1^*. The augmented matrix M_1 could not be an augmented minor of a Chinese postman problem $[M \mid b]$. However, its dual matrix M_1^* could be, and that matrix is forbidden for the Chinese postman problem to be an odd cut problem. But the graph associated with M_1^* as a Chinese postman problem is exactly the G_1^* associated with M_1 as an odd circuit problem.

Thus, we see that the forbidden augmented matrices are the duals of those in Theorem 5.2, but the graphs forbidden as augmented minors are the same. Theorem 5.3 is, thus, a duality result once the framework is understood.

THEOREM 5.4. *A co-postman problem is b-regular if, and only if, the associated graph does not contain any of the three augmented minors in Fig.* 7.

Proof. As for the Chinese postman problem, a co-postman problem is deletion regular. The augmented matrix $[M \mid b]$ can, therefore, only contain M_1^*, M_2^*, M_3^*. However, the graph we now draw is the dual because for the co-postman problem we draw the graph with respect to which the matrix is co-graphic.

THEOREM 5.5. *An odd circuit problem is b-regular if, and only if, the associated graph does not contain any of the three augmented minors* G_1, G_2, G_3.

6. Problems not co-postman. For $[M \mid b]$ a Chinese postman problem, it is easy to say when it is not a co-postman problem: if, and only if, the associated graph is not planar, i.e., contains a K_5 or a $K_{3,3}$ minor. By duality, the same answer applies to when an odd cut problem is an odd circuit problem. That is, let $[M^* \mid b^*]$ be an odd cut problem, let $[M \mid b]$ be its blocking Chinese postman problem, and let G be the augmented graph of the Chinese postman (and the odd cut) problem. Then, $[M \mid b]$ is also a co-postman problem if, and only if, G is planar, and then $[M \mid b]$ is the co-postman problem

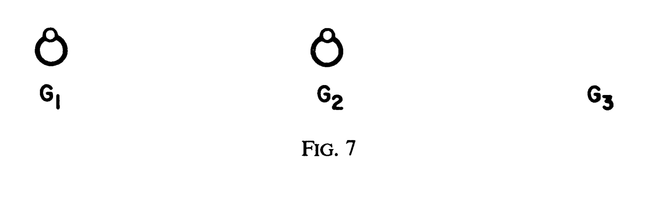


FIG. 7

on the dual augmented graph. Its blocking problem is, thus, an odd circuit problem on the dual augmented graph.

These two cases were relatively easy because we were considering the intersection of classes of problems that both involve deletion of b or both involve contraction of b. The other intersections require more work.

A problem $[M \mid b]$ is not a co-postman problem if, and only if, M is not co-graphic, i.e., M contains an F_7, F_7^*, K_5, or $K_{3,3}$ minor. Again we look for critical augmented minors of $[M \mid b]$. Clearly, $[M \mid b]$ must be b-regular, i.e., have no F_7 or F_7^* in M as considered in § 5.

Consider next K_5. The matrix M contains a K_5 minor means that $[M \mid b]$ contains as a minor either

even K_5

$$\left[\begin{array}{ccccccccccc|c}
1 & & & & & 1 & 1 & 1 & & & & 0 \\
 & 1 & & & & 1 & & & 1 & 1 & & 0 \\
 & & 1 & & & & 1 & & 1 & & 1 & 0 \\
 & & & 1 & & & & 1 & & 1 & 1 & 0 \\
 & & & & 1 & & & & & & & 1
\end{array}\right] \text{ or}$$

$$\left[\begin{array}{cccccccccc|c}
1 & & & & 1 & 1 & 1 & & & & b_1 \\
 & 1 & & & 1 & & & 1 & 1 & & b_2 \\
 & & 1 & & & 1 & & 1 & & 1 & b_3 \\
 & & & 1 & & & 1 & & 1 & 1 & b_4
\end{array}\right],$$

where not all of the b_i's are equal to zero.

For the second of these two augmented matrices, there is an associated augmented graph. The question is how many and which are not isomorphic. Here, isomorphic means simply changing the odd edges by interchanging on a circuit. Clearly, all $b = (b_1, b_2, b_3, b_4)$ with exactly one or exactly two b_i's equal to zero give isomorphic problems. Furthermore, all b with exactly three or all four b_i's equal to zero give isomorphic problems.

Consider next $K_{3,3}$. The matrix M contains a $K_{3,3}$, minor if, and only if $[M \mid b]$ contains as a minor one of the following two:

even $K_{3,3}$

$$\left[\begin{array}{cccccccccc|c}
1 & & & & & & 1 & 1 & 1 & 1 & 0 \\
 & 1 & & & & & 1 & 1 & & & 0 \\
 & & 1 & & & & 1 & & 1 & & 0 \\
 & & & 1 & & & & 1 & & 1 & 0 \\
 & & & & 1 & & & & 1 & 1 & 0 \\
 & & & & & 1 & & & & & 1
\end{array}\right]$$

$$\left[\begin{array}{ccccccccc|c}
1 & & & & & 1 & 1 & 1 & 1 & b_1 \\
 & 1 & & & & 1 & 1 & & & b_2 \\
 & & 1 & & & 1 & & 1 & & b_3 \\
 & & & 1 & & & 1 & & 1 & b_4 \\
 & & & & 1 & & & 1 & 1 & b_5
\end{array}\right],$$

where not all b_i's are equal to zero.

For the second of these two above augmented matrices, there is an associated augmented graph. There are five different right-hand sides b_i given in the following theorem.

THEOREM 6.1. *A problem* $[M \mid b]$ *is a co-postman problem if, and only if, it satisfies all of the following*:

(i) $[M \mid b]$ *is deletion regular*;

(ii) $[M \mid b]$ *contains no even* K_5 *or* $K_{3,3}$ *minor*;

(iii) $[M \mid b]$ *contains none of the following seven augmented minors* (*indicated by the different right-hand sides*):

$$\begin{array}{cccccccccc|cc} &&&&&&&&&& b^1 & b^2 \\ 1 &&&& 1 & 1 & 1 &&&& 0 & 1 \\ & 1 &&& 1 &&& 1 & 1 && 0 & 1 \\ && 1 &&& 1 && 1 && 1 & 0 & 1 \\ &&& 1 &&& 1 && 1 & 1 & 1 & 1 \end{array},$$

$$\begin{array}{ccccccccc|ccccc} &&&&&&&&& b^3 & b^4 & b^5 & b^6 & b^7 \\ 1 &&&&& 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ & 1 &&&& 1 & 1 &&& 0 & 1 & 0 & 1 & 1 \\ && 1 &&& 1 && 1 && 0 & 0 & 1 & 1 & 1 \\ &&& 1 &&& 1 && 1 & 0 & 0 & 1 & 1 & 1 \\ &&&& 1 &&& 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{array}.$$

Proof. From the previous remarks, we need only prove that the seven different b^i's are the only ones needed. Since M is either K_5 or $K_{3,3}$ the question is equivalent to the question: what are the different (not 2-isomorphic) Chinese postman problems on K_5 and on $K_{3,3}$? We give a lemma completing the proof.

LEMMA 6.2. *There are two different Chinese postman problems on* K_5: *one having two odd nodes and the other having four odd nodes. On* $K_{3,3}$, *there are five different Chinese postman problems with odd nodes as given in Fig.* 8.

Proof. There must be an even number of odd nodes, and no matter which even set of nodes is odd, we can renumber the nodes to be one of the augmented graphs given.

THEOREM 6.3. *An odd circuit problem is also a co-postman problem if, and only if, the associated augmented graph G of the odd circuit problem satisfies*

(i) *G is b-regular, i.e. contains no* G_1, G_2, G_3 *augmented minor*;

(ii) *G contains no even* K_5 *or* $K_{3,3}$ *minor*;

(iii) *G contains none of the four augmented minors in Fig.* 9.

Proof. Condition (i) of Theorem 6.1 implies that $[M \mid b]$ must be b-regular because it is contraction regular by being an odd circuit problem. By Theorem 5.5, an odd circuit problem is b-regular if, and only if, it contains no G_1, G_2, G_3 augmented minor.

Condition (ii) of Theorem 6.1 could occur because being an odd circuit problem only requires $[M \mid b]$ to be graphic, but not co-graphic, after contracting b.

Condition (iii) here is obtained from condition (iii) of Theorem 6.1 by drawing the augmented graphs for the matrices there with right-hand sides b^1, b^3, b^4, b^5. The rest of the proof consists of showing that the right-hand sides b^2, b^6, and b^7 need not be considered.

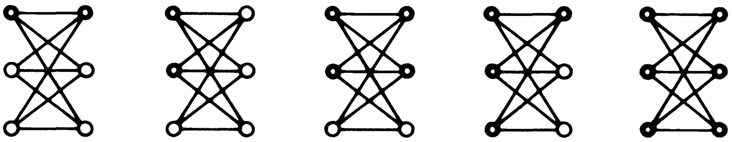


FIG. 8

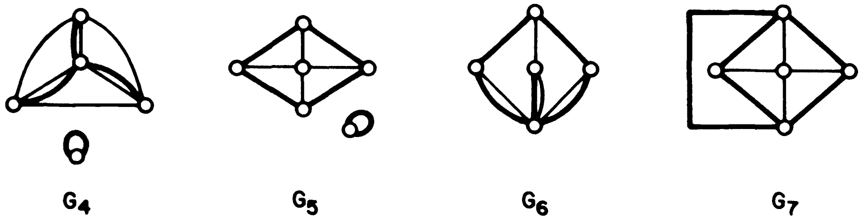


FIG. 9

Contracting b^2 gives an F_7 minor, contracting b^6 gives an F_7^* minor, and contracting b^7 gives a $K_{3,3}^*$ minor in their respective augmented matrices. Since $[M \mid b]$ is an odd circuit problem, it can have no F_7, F_7^*, $K_{3,3}^*$ minor after contracting b, completing the proof.

The augmented graph corresponding to an odd circuit problem is obtained by contracting b giving a graph G whose edges are even if the row deleted, after pivoting on b, had an entry equal to one. We draw the same graph for the odd circuit problem and the blocking co-postman problem. Thus, we have the following theorem.

THEOREM 6.4. *A co-postman problem is also an odd circuit problem if, and only if, the associated augmented graph G of the co-postman problem satisfies*

(i) *G is b-regular, i.e., contains no G_1, G_2, G_3 augmented minors*;

(ii) *G contains no even K_5 or $K_{3,3}$ minor*;

(iii) *G contains no G_4, G_5, G_6, or G_7 augmented minor.*

THEOREM 6.5. *An odd cut problem is also a co-postman problem if, and only if, the associated augmented graph G of the odd cut problem satisfies*:

(i) *G is b-regular, i.e. contains no G_1^*, G_2^*, G_3^* augmented minor*;

(ii) *G contains no odd $K_{3,3}$ minor*;

(iii) *G contains none of the four augmented minors in Fig.* 10.

Proof. The proof is similar to that of Theorem 6.3. Condition (iii) is as in Theorem 6.3 (ii), and the graphs listed are the augmented dual graphs there.

The reason that the even K_5 and $K_{3,3}$ need not be forbidden is that the corresponding augmented minor, e.g., for an even K_5 as a minor of an odd cut problem would be

$$\left[\begin{array}{ccccccccccc|c}
1 & & & & & & & 1 & 1 & & & 0 \\
 & 1 & & & & & & 1 & & 1 & & 0 \\
 & & 1 & & & & & 1 & & & 1 & 0 \\
 & & & 1 & & & & & 1 & 1 & & 0 \\
 & & & & 1 & & & & 1 & & 1 & 0 \\
 & & & & & 1 & & & & 1 & 1 & 0 \\
 & & & & & & 1 & & & & & 1
\end{array}\right]$$

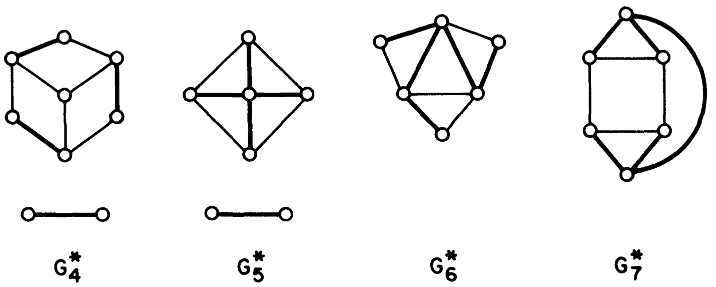


FIG. 10

and that augmented matrix is not forbidden for a co-postman problem because it is cographic after deletion of the right-hand side.

The odd $K_{3,3}$ minor arises from b^7 as a right-hand side in Theorem 6.1(iii). That case was excluded for an odd circuit problem but cannot be excluded for an odd cut problem. However, b^2 and b^6 can be excluded as before. The odd K_5 minor does not arise but is, in fact, excluded by G_3^*, which is an augmented minor of an odd K_5 minor. That is, an odd K_5 minor is not b-regular, but an odd $K_{3,3}$ minor is b-regular.

THEOREM 6.6. *A Chinese postman problem is also an odd circuit problem if, and only if, the associated augmented graph G of the Chinese postman problem satisfies*

(i) *G is b regular, i.e. contains no G_1^*, G_2^*, G_3^* augmented minor*;

(ii) *G contains no odd $K_{3,3}$ minor*;

(iii) *G contains none of the four augmented minors G_4^*, G_5^*, G_6^*, G_7^*.*

7. Problems not Chinese postman.

THEOREM 7.1. *A problem $[M \mid b]$ is a Chinese postman problem if, and only if, it satisfies all of the following*:

(i) *$[M \mid b]$ is deletion regular*;

(ii) *$[M \mid b]$ contains no even K_5^* or $K_{3,3}^*$ minor*;

(iii) *$[M \mid b]$ contains none of the following augmented minors indicated by the different right-hand sides*

$$\begin{array}{cccccccccc|cccccc}
 & & & & & & & & & & b^1 & b^2 & b^3 & b^4 & b^5 & b^6 \\
1 & & & & & & 1 & 1 & & & 0 & 0 & 1 & 1 & 0 & 1 \\
 & 1 & & & & & 1 & & 1 & & 0 & 0 & 0 & 0 & 0 & 1 \\
 & & 1 & & & & 1 & & & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
 & & & 1 & & & & 1 & 1 & & 0 & 0 & 0 & 0 & 1 & 1 \\
 & & & & 1 & & & 1 & & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\
 & & & & & 1 & & & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array},$$

$$\begin{array}{ccccccccc|cc}
 & & & & & & & & & b^7 & b^8 \\
1 & & & & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\
 & 1 & & & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
 & & 1 & & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
 & & & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1
\end{array}.$$

Proof. The proof is similar to that of Lemma 6.2 except for the proof that the right-hand sides given in (iii) suffice. That proof is provided by Lemmas 7.2 and 7.3.

LEMMA 7.2. *There are six different co-postman problems on K_5 given by the augmented graphs in Fig.* 11.

Proof. The proof is given by Fig. 12. In that figure, we start with one of the six augmented graphs in Fig. 11 and show what cuts to interchange to get the next graph. In Fig. 12, the six augmented graphs are across the bottom, and the changes are from bottom to top.

LEMMA 7.3. *There are two different co-postman problems on $K_{3,3}$ given by the augmented graphs in Fig.* 13.

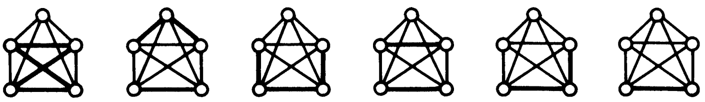


FIG. 11

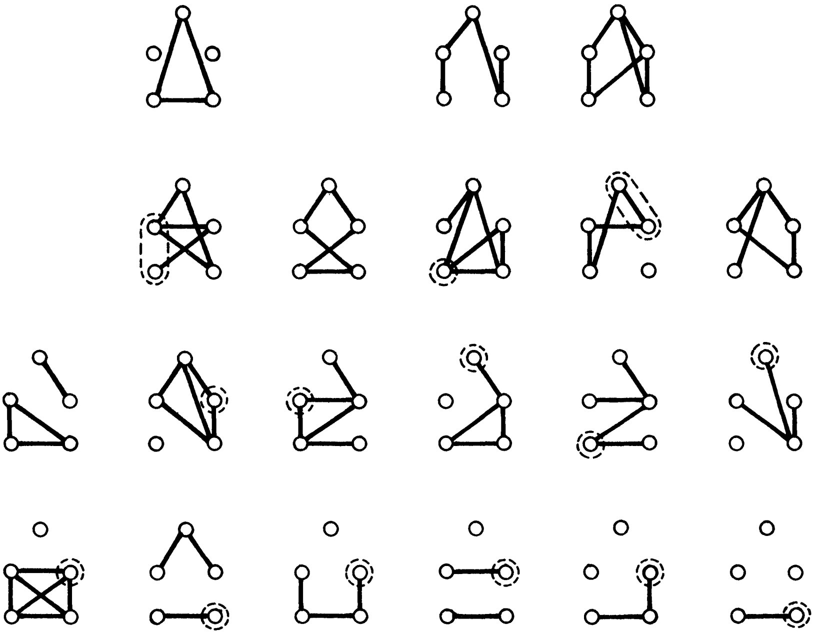


FIG. 12

Proof. The proof is similar to that of Lemma 7.2 and is given by Fig. 14. The two different problems are on the left and changes are made from left to right. The assertion here is that every co-independent subset of edges of $K_{3,3}$ is present here.

THEOREM 7.4. *An odd circuit problem is also a Chinese postman problem if, and only if, the associated augmented graph G of the odd circuit problem satisfies*

(i) *G is b-regular, i.e., contains no G_1, G_2, G_3 augmented minor;*

(ii) *G contains none of the six augmented minors G_8, G_9, G_{10}, G_{11}, G_{12}, G_{13} given in Fig.* 15.

Proof. An even K_5^* or $K_{3,3}^*$ could not be present in an odd circuit problem so need not be excluded.

The right-hand side b^1 in Theorem 7.1(iii) gives G_8; b^2 gives G_9; b_3 gives G_{13}; b^4 gives G_{10}; b^5, after contraction, gives a $K_{3,3}^*$ minor, and so does b^6. The right-hand sides b^7 and b^8 give G_{11} and G_{12}. The reason that G_{13} is listed last is that it is the only one that is not planar.

THEOREM 7.5. *A co-postman problem is also an odd cut problem if, and only if, the associated augmented graph G of the co-postman problem satisfies*

(i) *G is b-regular, i.e., contains no G_1, G_2, G_3 augmented minor;*

(ii) *G contains none of the six augmented minors: G_8, G_9, G_{10}, G_{11}, G_{12}, or G_{13}.*

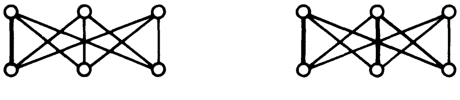


FIG. 13

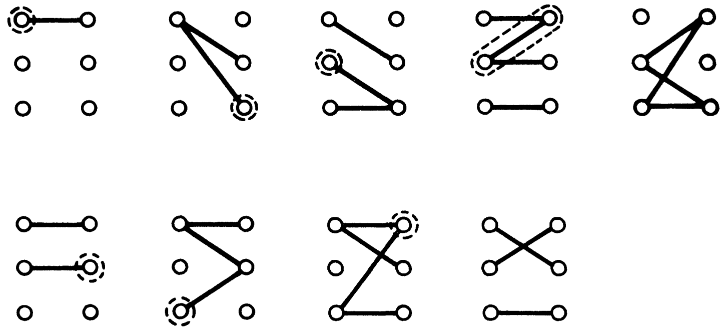


FIG. 14

THEOREM 7.6. *An odd cut problem is also a Chinese postman problem if, and only if, the associated augmented graph G of the odd cut problem satisfies*

(i) *G is b-regular, i.e. contains no* G_1^*, G_2^*, G_3^* *augumented minor*;

(ii) *G contains no even* K_5 *or* $K_{3,3}$ *minor*;

(iii) *G contains none of the six augmented minors* G_8^*, G_9^*, G_{10}^*, G_{11}^*, G_{12}^*, G_{14} *in Fig.* 16.

THEOREM 7.7. *A Chinese postman problem is also an odd cut problem if, and only if, the associated augmented graph for the Chinese postman problem satisfies*

(i) *G is b-regular, i.e. contains no* G_1^*, G_2^*, G_3^* *augmented minor*;

(ii) *G contains no even* K_5 *or* $K_{3,3}$ *minor*;

(iii) *G contains none of the six augmented minors* G_8^*, G_9^*, G_{10}^*, G_{11}^*, G_{12}^*, G_{14} *in Fig.* 16.

Proof. The proof is similar to that of the previous theorems. Here, contracting b^1 in Theorem 7.1 (iii) gives a matrix which is co-graphic with respect to G_8^*. Similarly, b^2 gives G_9^*, b^4 gives G_{10}^*, b^5 gives G_{14}, b^7 gives G_{11}^*, and b^8 gives G_{12}^*. The other right-hand sides give augmented matrices that could not be odd cut problems because contracting b^3 or b^6 gives a $K_{3,3}$ minor.

8. Some special cases and examples. Define a graph to be *outer planar* if it can be drawn so that every node is on the outside of the graph. A graph is outer planar if, and only if, it has no K_4 or $K_{2,3}$ minor [6]. Define a graph to be *inner planar* if it has no $K_4^*(=K_4)$ or $K_{2,3}^*$ minor. By duality, a graph is inner planar if, and only if, it can be drawn so that some one node is in every region.

THEOREM 8.1. *A Chinese postman problem* $[M \mid b]$ *on a graph G that is outer planar is also a co-postman problem and an odd cut problem.*

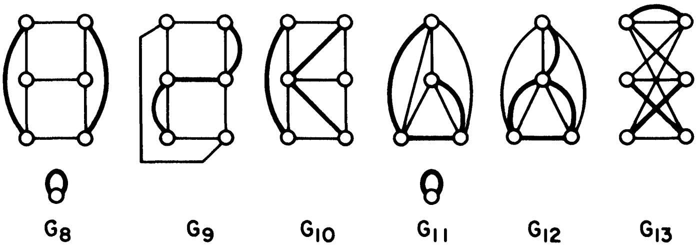


FIG. 15

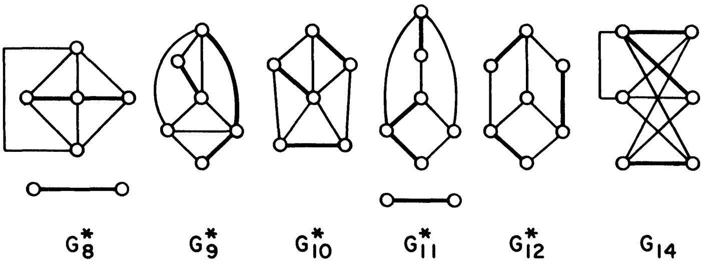


FIG. 16

Proof. Since the graph must be planar, $[M \mid b]$ is also a co-postman problem. It is an odd cut problem because all of the forbidden augmented minors in Theorem 7.7 include K_4 or $K_{2,3}$ minors.

The Chinese postman problem on G_{10}^* is an example of the theorem. The co-postman problem over the same matrix was given in Fig. 5(b). As discussed in § 4, the same matrix is an odd cut problem on $K_{3,3}$, which is not planar, so the problem is not an odd circuit problem.

THEOREM 8.2. *A co-postman problem on a graph G that is inner planar is also a Chinese postman problem and an odd circuit problem.*

There are corresponding results from duality for odd cut and odd circuit problems: an odd cut problem on an outer planar graph is also an odd circuit problem and a Chinese postman problem; and an odd circuit problem on an inner planar graph is also an odd cut problem and is a co-postman problem. These results apply regardless of the right-hand side.

We remark that the problems given as forbidden minors provide the following examples of intersections of problems: only Chinese postman—K_5 with 4 odd nodes; Chinese postman and co-postman—M_1^*, M_2^*, M_3^*; Chinese postman and odd cut—odd $K_{3,3}$; Chinese postman and odd circuit—M_{14}^*, even K_5; Chinese postman, co-postman, and odd cut—$M_8^*, \cdots, M_{12}^*$; Chinese postman, co-postman, and odd circuit—$M_4^*, \cdots,$ M_7^*; co-postman and odd circuit—M_{13}; only co-postman—K_5 with all odd edges. The other intersections are blocking to one of these except for all four problems for which an example was given in § 4.

The problem M_{13} is neither Chinese postman nor odd cut yet does have the Fulkerson property. The facets of the associated polyhedron have been explicitly calculated and verified to be equal to the facets given by the appropriate clutter. Thus, we have a counterexample to the conjecture that every binary group problem satisfying the Fulkerson property is either a Chinese postman problem or an odd cut problem. We conjecture that the only problems among these problems for which the Fulkerson property does not hold are the problems that are only co-postman or only odd circuit. Since the Fulkerson property is known to hold for Chinese postman and odd cut problems, this conjecture is equivalent to saying that the Fulkerson property holds for problems that are both odd circuit and co-postman problems.

There are only three known critical cases where the Fulkerson property does not hold [3], [14]. One involves F_7^* so is not among the problems considered here, and the other two are the blocking pair: the co-postman and odd circuit problems on K_5 with all edges odd (the first case in Fig. 12). There are six different odd circuit (or co-postman) problems on K_5 (Fig. 11). For these six odd circuit problems, the six right-hand sides give a Chinese postman problem (on G_8^*, G_9^*, G_{10}^*, and G_{14}) for four different right-hand sides; one right-hand side gives the augmented matrix M_{13} that has the Fulkerson property and is both co-postman and odd circuit (on G_{13} and K_5, respectively); and the sixth right-hand side gives R_{10} when deleted, and this problem does not have the Fulkerson property and is only an odd circuit problem. Thus, no K_5 minor is sufficient for the Fulkerson property to hold but seems to be far from necessary.

Seymour's results [14] on matroids having the max-flow min-cut property show that a Chinese postman problem has that property if and only if it has no odd K_4 minor (K_4 with all odd nodes). The forbidden augmented minor for the odd cut problem is $K_{2,3}$ with four odd nodes including all of the degree two nodes. For the odd circuit problem, the forbidden minor is $K_{2,3}^*$, i.e. a doubled triangle. It is interesting to note how frequently those minors occur as augmented minors of our forbidden minors here. Both the Chinese postman and its blocking odd cut problem have the max-flow min-cut property on the outer planar graphs previously discussed.

REFERENCES

[1] J. EDMONDS AND E. L. JOHNSON, *Matchings, Euler tours, and the Chinese postman*, Math. Programming, 5 (1973), pp. 88–124.

[2] D. R. FULKERSON, *Blocking polyhedra*, in Graph Theory and its Applications, B. Harris, ed., Academic Press, New York, 1970, pp. 93–112.

[3] G. GASTOU AND E. L. JOHNSON, *Binary group and Chinese postman polyhedra*, Math. Programming, 34 (1986), pp. 1–33.

[4] R. E. GOMORY, *Some polyhedra related to combinatorial problems*, Linear Algebra Appl., 2 (1969), pp. 451–558.

[5] M. GROETSCHEL AND W. R. PULLEYBLANK, *Weakly bipartite graphs*, Oper. Res. Lett., 7 (1981), pp. 23–27.

[6] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[7] A. LEHMAN, *A solution of the Shannon switching game*, SIAM J. Appl. Math., 12 (1964), pp. 687–725.

[8] ———, *Matroids and ports*, Notices Amer. Math. Soc., 12 (1965), p. 342.

[9] ———, *On the width-length inequality*, Math. Programming, 17 (1979), pp. 403–417.

[10] MEI-KO KWAN, *Graphic programming using odd or even points*, Chinese J. Math., 1 (1962), pp. 273–277.

[11] S. MOSTERTS, *Algebraic polyhedra and polyhedral aspects of binary group problems*, Tesi di laurea, University of Pisa, 1983.

[12] M. W. PADBERG AND M. R. RAO, *Odd minimum cut sets and b-matchings*, Math. Oper. Res., 7 (1982), pp. 67–80.

[13] P. D. SEYMOUR, *The forbidden minors of binary clutters*, J. London Math. Soc. (2), 12 (1976), pp. 356–360.

[14] ———, *Matroids with the max-flow min-cut property*, J. Combin. Theory Ser. B, 23 (1977), pp. 189–222.

[15] W. T. TUTTE, *An algorithm for determining whether a given binary matroid is graphic*, Proc. Amer. Math. Soc., 11 (1960), pp. 903–917.

[16] ———, *Introduction to the Theory of Matroids*, American Elsevier, New York, 1971.

[17] H. WHITNEY, *2-isomorphic graphs*, Amer. J. Math., 55 (1933), pp. 245–254.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

QUADRATIC CONES INVARIANT UNDER SOME LINEAR OPERATORS*

DRAGOMIR Ž. ĐOKOVIĆ†

Abstract. A (solid) quadratic cone K in a finite-dimensional vector space V (over **R**, **C**, or **H**) is the set of all $x \in V$ satisfying $f(x, x) \geqq 0$, where f is a fixed indefinite hermitian form. Given such a cone K, we characterize the linear operators A for which $AK \subset K$, and also those for which $AK = K$. We also show that if $\rho(A) = \nu(A)$ for some (multiplicative) norm ν on the algebra of linear operators (ρ denotes the spectral radius) then there exists an A-invariant quadratic cone of specified signature. For this purpose we strengthen a result of Mott and Schneider characterizing the operators A for which $\rho(A) = \nu(A)$ is possible.

Key words. spectral radius, multiplicative norm, semisimple operator, Jordan decomposition, ice-cream cone, hermitian form, real quaternions

AMS(MOS) subject classifications. 15A48, 15A60, 15A63

1. Introduction. If A is a linear operator on a finite-dimensional complex vector space V and ν is a (multiplicative) norm on the algebra $\mathscr{A}$ of operators on V then it is well known that $\rho(A) \leqq \nu(A)$, where ρ denotes the spectral radius. Mott and Schneider [8], see also [5, § 2.3], have shown that

$$\rho(A) = \inf_{\nu} \nu(A).$$

Furthermore they have shown that there exists a ν such that $\rho(A) = \nu(A)$ iff every eigenvalue λ of A with $|\lambda| = \rho(A)$ is a simple root of the minimal polynomial of A.

The above results remain valid (Theorem 5) when the class of all (multiplicative) norms on $\mathscr{A}$ is replaced by the class of norms on $\mathscr{A}$ induced by inner products on V; such norms we call Hilbert norms. We also show that the same results are valid for real and quaternionic spaces.

The problem of characterizing linear operators A on $\mathbf{R}^n$ which leave invariant a fixed proper cone K (or alternatively, which leave invariant a proper cone K belonging to a specified class of cones) has been extensively studied, see for instance [7], [10], [11] and the references mentioned therein. The class of quadratic cones K in $V = D^n$ ($D = \mathbf{R}$, $\mathbf{C}$, or $\mathbf{H}$) is of special interest. Such a cone consists of all $x \in V$ satisfying $f(x, x) \geqq 0$ where f is an indefinite hermitian form on V. Our Theorem 7 shows that two theorems of Loewy and Schneider [7, Thms. 2.3 and 2.4] generalize to arbitrary quadratic cones.

Theorem 7 admits a nice geometric interpretation. Namely let P be the projective space attached to V and let S be the hermitian hyperquadric in P defined by the equation $f(x, x) = 0$ (f is an indefinite hermitian form). The complement of S in P has two connected components and let P^+ be the closure of one of them. Then Theorem 7 gives a characterization of those projective transformations of P which leave P^+ invariant, and of those which map P^+ onto itself.

As an application of Theorem 5 we show that if $\rho(A) = \nu(A)$ for some norm ν and if $\rho(A)$ is an eigenvalue of A then A leaves invariant a quadratic cone of signature $(n - 1, 0, 1)$.

We conclude with an improvement of a result of Vandergraft [11] characterizing operators A, in the case $D = \mathbf{R}$, for which there exists a proper cone K such that A maps $K - \{0\}$ in the interior of K. We show that his characterization remains valid when K is restricted to be an ice-cream cone.

* Received by the editors March 8, 1983; accepted for publication (in revised form) May 27, 1986.

† Department of Pure Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

2. Notation and terminology. We denote by D one of the three classical fields $\mathbf{R}$, $\mathbf{C}$, or $\mathbf{H}$ and by F the center of D. By $\mathscr{A}$ we denote the F-algebra of $n \times n$ matrices over D. When convenient we shall view $\mathscr{A}$ as the algebra of linear operators of the right D-vector space $V = D^n$. The elements of V are viewed as column vectors.

For $A \in \mathscr{A}$ and $\lambda \in \mathbf{C}$ we define $A_\lambda \in \mathscr{A}$ by

$$A_\lambda = \begin{cases} A - \lambda I_n & \text{if } \lambda \in \mathbf{R} \text{ or } D = \mathbf{C}, \\ A^2 - (\lambda + \bar{\lambda})A + |\lambda|^2 I_n & \text{otherwise} \end{cases}$$

where I_n is the identity matrix.

For $A \in \mathscr{A}$ we define its *spectrum* $\sigma(A) \subset \mathbf{C}$ by

$$\sigma(A) = \{\lambda \in \mathbf{C} : A_\lambda \text{ is singular}\}.$$

Observe that $\sigma(A)$ is a finite subset of $\mathbf{C}$ and its cardinality is at most n if $D = \mathbf{R}$ or $\mathbf{C}$ and at most $2n$ if $D = \mathbf{H}$.

The *spectral radius* $\rho(A)$ of A is defined by

$$\rho(A) = \sup \{|\lambda| : \lambda \in \sigma(A)\}.$$

We say that A is *semisimple at* $\lambda \in \mathbf{C}$ if A_λ and $(A_\lambda)^2$ have the same rank. Clearly A is semisimple at λ if $\lambda \notin \sigma(A)$. If A is semisimple at λ, $\forall \lambda \in \mathbf{C}$, then A is said to be *semisimple*. It is well known that every $A \in \mathscr{A}$ can be written uniquely as $A = S + N$ where $SN = NS$, S is semisimple, and N is nilpotent. This is known as the *Jordan decomposition* of A.

A *norm* on $\mathscr{A}$ is a map $\nu : \mathscr{A} \to \mathbf{R}$ satisfying the conditions

1° $\nu(A) \geqq 0, \quad \forall A \in \mathscr{A},$

2° $\nu(A) = 0 \Leftrightarrow A = 0,$

3° $\nu(\lambda A) = |\lambda| \nu(A) \quad \forall A \in \mathscr{A}, \quad \forall \lambda \in F,$

4° $\nu(A + B) \leqq \nu(A) + \nu(B) \quad \forall A, B \in \mathscr{A},$

5° $\nu(AB) \leqq \nu(A)\nu(B) \quad \forall A, B \in \mathscr{A}.$

Let $f : V \times V \to D$ be an inner product on V, i.e., f is a positive definite hermitian form. Then (V, f) is a finite-dimensional Hilbert space over D. The norm in V induced by f will be denoted by $|\cdot|_f$; thus $|x|_f^2 = f(x, x)$, $\forall x \in V$. The norm $|\cdot|_f$ on V induces a norm ν_f on $\mathscr{A}$ which is defined by

$$\nu_f(A) = \sup \left\{ \frac{|Ax|_f}{|x|_f} : x \in V - \{0\} \right\}.$$

Such norms ν_f will be called *Hilbert norms.*

3. Norm and spectral radius. Let ν be a norm on $\mathscr{A}$ and let U be the closed unit ball in $\mathscr{A}$, i.e.,

$$U = \{X \in \mathscr{A} : \nu(X) \leqq 1\}. \tag{3.1}$$

It is well known that U is a compact subset of $\mathscr{A}$. Lemma 1 is a simple consequence of the compactness of U.

LEMMA 1. *For $A \in \mathscr{A}$ we have $\rho(A) \leqq \nu(A)$. Furthermore if $\rho(A) = \nu(A)$ then A is semisimple at every λ such that $|\lambda| = \rho(A)$.*

Proof. The case $D = \mathbf{R}$ (resp. $D = \mathbf{H}$) can be reduced to the case $D = \mathbf{C}$ by complexification (resp. restriction of scalars). Thus we assume that $D = \mathbf{C}$. Then the first assertion was proved by Mott and Schneider [8, Thm. 2].

To prove the second assertion, we may assume (without any loss of generality) that $\nu(A) = \rho(A) = 1$. Let $\lambda \in \sigma(A)$ with $|\lambda| = 1$. Assume that A is not semisimple at λ. Then there exists linearly independent vectors $v, w \in V$ such that $Av = v\lambda$ and $Aw = w\lambda + v$. It follows that $A^k w = w\lambda^k + \lambda^{k-1}kv$ for each $k \geqq 0$. Since $A^k w \in Uw$ and Uw is compact, the sequence $w\lambda^k + \lambda^{k-1}kv$ must be bounded. Since $|\lambda| = 1$, this implies that $v = 0$, which is a contradiction. Hence A must be semisimple at λ.

LEMMA 2. *Let $A = S + N$ be the Jordan decomposition of $A \in \mathscr{A}$. If $\varepsilon \in F$, $\varepsilon \neq 0$, then $S + \varepsilon N$ is similar to A.*

Proof. If $D = \mathbf{C}$ or $\mathbf{H}$ this follows immediately from the existence of the canonical Jordan form. The case $D = \mathbf{R}$ reduces to the case $D = \mathbf{C}$ via complexification.

LEMMA 3. *Let g be an inner product in V, $T \in \mathscr{A}$ a nonsingular matrix and f a new inner product defined by $f(x, y) = g(Tx, Ty)$. Then $\nu_f(A) = \nu_g(TAT^{-1})$, $\forall A \in \mathscr{A}$.*

Proof. This follows from the fact that $|x|_f = |Tx|_g$, $\forall x \in V$.

Mott and Schneider [8, Thm. 1] have shown that (in the case $D = \mathbf{C}$) for all $A \in \mathscr{A}$ we have $\rho(A) = \inf \nu(A)$ where inf is taken over all norms on $\mathscr{A}$. A similar statement is valid when inf is taken over all transform absolute norms. This stronger result is stated, without proof in [3], and is attributed to Saunders and Schneider [9]. The next lemma is a consequence of this stronger result. For the convenience of the reader we shall include its proof.

LEMMA 4. *For $A \in \mathscr{A}$ we have $\rho(A) = \inf \nu_f(A)$ where* inf *is taken over all inner products f on V.*

Proof. In view of Lemma 1 it suffices to show that if $r > \rho(A)$ then there exists an inner product f on V such that $\nu_f(A) < r$. Let $A = S + N$ be the Jordan decomposition of A (with S semisimple and N nilpotent). There exists an inner product g on V such that S is a normal operator of the Hilbert space (V, g). Hence

$$\nu_g(S) = \rho(S) = \rho(A) < r.$$

Choose $\varepsilon > 0$ small enough so that $\nu_g(S + \varepsilon N) < r$. By Lemma 2 there exists an invertible matrix $T \in \mathscr{A}$ such that $TAT^{-1} = S + \varepsilon N$. Define a new inner product f on V by $f(x, y) = g(Tx, Ty)$. Then by Lemma 3 we have

$$\nu_f(A) = \nu_g(TAT^{-1}) = \nu_g(S + \varepsilon N) < r. \qquad \text{QED}$$

In the case $D = \mathbf{C}$, Mott and Schneider [8, Thm. 2] have characterized the operators $A \in \mathscr{A}$ for which there exists a norm ν on $\mathscr{A}$ such that $\nu(A) = \rho(A)$. This is contained as the part (ii) $\Leftrightarrow$ (iii) of the next theorem. This theorem is probably not new but for the lack of reference we shall include a proof.

THEOREM 5. *For $A \in \mathscr{A}$ the following are equivalent:*

(i) $\exists$ *inner product f on V such that $\nu_f(A) = \rho(A)$;*

(ii) $\exists$ *norm ν on $\mathscr{A}$ such that $\nu(A) = \rho(A)$;*

(iii) *A is semisimple at every $\lambda \in \sigma(A)$ with $|\lambda| = \rho(A)$.*

Proof. (i) $\Rightarrow$ (ii) is trivial.

(ii) $\Rightarrow$ (iii) is contained in Lemma 1.

(iii) $\Rightarrow$ (i) We may assume that $A \neq 0$. Then (iii) implies that $\rho(A) > 0$. Replacing A by $\rho(A)^{-1}A$, we may assume that $\rho(A) = 1$. There is a unique decomposition $V = V_1 \oplus V_2$ into A-invariant subspaces such that the restrictions $A_k = A|V_k$ $(k = 1, 2)$ satisfy:

$$\sigma(A_1) \subset \{\lambda \in \mathbf{C} : |\lambda| = 1\},$$

$$\sigma(A_2) \subset \{\lambda \in \mathbf{C} : |\lambda| < 1\}.$$

In view of Lemma 4 and Lemma 6 (below) it suffices to consider the case $V = V_1$. Then (iii) implies that A is semisimple and consequently there exists an inner product

$f\colon V \times V \to D$ such that A is a unitary operator of the Hilbert space (V, f). In particular $\nu_f(A) = 1$.

This concludes the proof of the theorem.

LEMMA 6. *Let $A \in \mathscr{A}$ and let $V = V_1 \oplus V_2$ be a decomposition into A-invariant subspaces. Let f be an inner product on V such that $V_1 \perp V_2$ and let f_k (resp. A_k) be the restriction of f (resp. A) to $V_k \times V_k$ (resp. V_k). Then writing $\nu_k = \nu_{f_k}$ $(k = 1, 2)$ we have*

$$\nu_f(A) = \max(\nu_1(A_1), \nu_2(A_2)).$$

Proof. Observe first that if α, β, γ, $\delta > 0$ then

$$\frac{\alpha+\beta}{\gamma+\delta} \leqq \max\left(\frac{\alpha}{\gamma}, \frac{\beta}{\delta}\right).$$

This implies that (we write $|x|$ instead of $|x|_f$)

$$\sup_{(x_1, x_2) \neq (0,0)} \frac{|A_1x_1|^2 + |A_2x_2|^2}{|x_1|^2 + |x_2|^2} \leqq \max(\nu_1(A_1)^2, \nu_2(A_2)^2),$$

where $x_k \in V_k$ $(k = 1, 2)$. For $x \in V$ we can write $x = x_1 + x_2$ $(x_k \in V_k)$ and so $Ax = A_1x_1 + A_2x_2$ and

$$|Ax|^2 = |A_1x_1|^2 + |A_2x_2|^2, \qquad |x|^2 = |x_1|^2 + |x_2|^2.$$

Hence the above inequality implies that $\nu_f(A) \leqq \max(\nu_1(A_1), \nu_2(A_2))$. On the other hand the inequalities $\nu_f(A) \geqq \nu_k(A_k)$ $(k = 1, 2)$ are obvious.

Remark 1. Lemma 4 and the equivalence (i) $\Leftrightarrow$ (iii) of Theorem 5 as well as their proofs are reminiscent of two theorems of W. Givens [4]. For $A \in \mathscr{A}$, in the case $D = \mathbf{C}$, and for an inner product f on V the *field of values* of A is defined as

$$\Phi_f(A) = \{f(x, Ax) : |x|_f = 1\}.$$

It is well known that $\sigma(A) \subset \Phi_f(A)$. Givens showed that if $\Delta(A)$ is the convex hull of $\sigma(A)$ then $\Delta(A) = \bigcap \Phi_f(A)$, where the intersection is over all inner products f on V. He also showed that there exists an inner product f such that $\Delta(A) = \Phi_f(A)$ iff A is semisimple at every λ on the boundary of $\Delta(A)$.

4. Invariant Quadratic Cones. Let $f\colon V \times V \to D$ be a hermitian form. Its *signature* sign (f) is the triple (n_-, n_0, n_+) where n_- (resp. n_+) is the maximum dimension (over D) of a subspace of V on which the restriction of f is negative definite (resp. positive definite) and $n_- + n_0 + n_+ = n$. Thus $n_- + n_+ = r$ is the rank of f, and n_0 is the dimension of its radical Rad f. We say that f is *indefinite* if $n_- \geqq 1$ and $n_+ \geqq 1$. If $n_- = 0$ (resp. $n_+ = 0$) we write $f \geqq 0$ (resp. $f \leqq 0$).

A (solid) *quadratic cone* in V is a subset

$$V_f^+ = \{x \in V : f(x, x) \geqq 0\}$$

where f is an indefinite hermitian form on V. We say that this cone has *type* (n_-, n_0, n_+) if sign $(f) = (n_-, n_0, n_+)$.

We raise two problems about linear operators preserving quadratic cones.

PROBLEM 1. Characterize linear operators $A \in \mathscr{A}$ which leave invariant a fixed quadratic cone K in V.

PROBLEM 2. Characterize linear operators $A \in \mathscr{A}$ which leave invariant some quadratic cone K in V of fixed type (n_-, n_0, n_+).

In the case $D = \mathbf{R}$ the first problem was attacked by Loewy and Schneider [7] for the cones of type $(n - 1, 0, 1)$. Some of their results are generalized in the following:

THEOREM 7. *Let f be an indefinite hermitian form on V, $A \in \mathscr{A}$, and let g be the hermitian form on V defined by $g(x, y) = f(Ax, Ay)$. Then for the statements*

(i) $A \cdot V_f^+ \subset V_f^+$,

(ii) $\exists t \geqq 0, \quad g - tf \geqq 0$,

(iii) $A \cdot V_f^+ = V_f^+$,

(iv) $\exists t > 0, \quad g = tf$,

we have (i) $\Leftrightarrow$ (ii) *and* (iii) $\Rightarrow$ (iv). *If A is invertible then* (iv) $\Rightarrow$ (iii).

Proof. (i) $\Rightarrow$ (ii) From (i) we have $f(x, x) \geqq 0 \Rightarrow g(x, x) \geqq 0$. Hence (ii) follows from [2, Thm. 5].

(ii) $\Rightarrow$ (i) If $x \in V_f^+$ then (ii) implies that $f(Ax, Ax) = g(x, x) \geqq tf(x, x) \geqq 0$, i.e., $Ax \in V_f^+$.

(iii) $\Rightarrow$ (iv) It follows from (iii) that A is invertible and that it maps the boundary Δ of V_f^+ onto itself. Thus $f(x, x) = 0 \Rightarrow g(x, x) = 0$. By applying a result of Krein and Šmul'jan [6] (see also [2, Thm. 3]) we conclude that $g = tf$ for some $t \in \mathbf{R}$. Since f is indefinite, (iii) implies that $t > 0$.

The proof of the last assertion of the theorem is straightforward.

Remark 2. If H is the matrix of f (with respect to the standard basis of $V = D^n$) then (ii) and (iv) can be written as: $\exists t \geqq 0$, $A^*HA - tH \geqq 0$, and $\exists t > 0$, $A^*HA = tH$, respectively, where A^* denotes the conjugate transpose of A.

Our next theorem deals with Problem 2. By K^0 we denote the interior of the cone K.

THEOREM 8. *Let $n \geqq 2$ and $A \in \mathscr{A}$. Then we have the following*:

(i) *Assume that $\rho(A) \in \sigma(A)$ and that A is semisimple at every λ with $|\lambda| = \rho(A)$. Then A leaves invariant a quadratic cone K of type* $(n - 1, 0, 1)$.

(ii) *Assume that $\rho(A)$ is a simple eigenvalue of A and that $|\lambda| < \rho(A)$ for all other $\lambda \in \sigma(A)$. Then there exists a quadratic cone K of type* $(n - 1, 0, 1)$ *such that* $x \in K - \{0\} \Rightarrow Ax \in K^0$.

Proof. We shall prove (i) and (ii) simultaneously. Let $a \in V$, $a \neq 0$, satisfy $Aa = a \cdot \rho(A)$. The hypotheses imply that there exists an A-invariant hyperplane W such that $a \notin W$. If B is the restriction of A to W then $\rho(B) \leqq \rho(A)$ with strict inequality in case (ii).

If $\rho(A) = 0$, which is possible only in case (i), then $A = 0$ and (i) trivially holds. Thus we may assume that $\rho(A) > 0$. Multiplying A by a suitable positive scalar we may assume that

$$\rho(B) \leqq 1 \leqq \rho(A) \tag{4.1}$$

with strict second inequality in case (ii).

By Theorem 5 there exists an inner product g on W such that $\rho(B) \leqq \nu_g(B) \leqq 1$. Hence for $x \in W$ we have

$$|x|_g \leqq 1 \Rightarrow |Bx|_g \leqq 1. \tag{4.2}$$

For $x, x' \in V$ we can write $x = y + at$, $x' = y' + at'$ where $y, y' \in W$ and $t, t' \in D$. We define a hermitian form f on V by

$$f(x, x') = f(y + at, y' + at') = \bar{t}t' - g(y, y'). \tag{4.3}$$

It is clear that sign $(f) = (n - 1, 0, 1)$ and so $K = V_f^+$ is a quadratic cone of type $(n - 1, 0, 1)$.

Let $x = y + at \in K - \{0\}$, $y \in W$, $t \in D$. It follows from (4.3) that $|t| \geqq |y|_g$ and so $t \neq 0$. Using (4.1) and (4.2) we obtain that

$$|t|^2\rho(A)^2 \geqq |t|^2 \geqq |y|_g^2 \geqq |By|_g^2. \tag{4.4}$$

Since $Ax = A(y + at) = By + a\rho(A)t$, it follows from (4.4) and (4.3) that $f(Ax, Ax) \geqq 0$, i.e., $Ax \in K$. Thus A leaves K invariant. In the case (ii) the first inequality in (4.4) is strict and consequently $Ax \in K^0$. This completes the proof.

Remark 3. Let $D = \mathbf{R}$ and let K be a quadratic cone of type $(n - 1, 0, 1)$. Then K^0 has two connected components. If K_1^0 is one of these components then its closure K_1 will be called an ice-cream cone. Furthermore K_1 is a proper cone, i.e., it is a closed convex cone with nonempty interior satisfying $K_1 \cap (-K_1) = \{0\}$. It is easy to see that both assertions of Theorem 8 remain valid for ice-cream cones.

THEOREM 9. *When $D = \mathbf{R}$ then for $A \in \mathscr{A}$ the following are equivalent*:

(i) *$\rho(A)$ is a simple eigenvalue of A, greater than the magnitude of any other eigenvalue*;

(ii) *$\exists$ proper cone K such that $x \in K - \{0\} \Rightarrow Ax \in K^0$*;

(iii) *$\exists$ ice-cream cone K such that $x \in K - \{0\} \Rightarrow Ax \in K^0$.*

Proof. (i) $\Leftrightarrow$ (ii) is a result of Vandergraft [11, Thm. 4.4].

(i) $\Rightarrow$ (iii) follows from Theorem 8 (ii) and Remark 3.

(iii) $\Rightarrow$ (ii) is trivial.

Acknowledgments. I would like to thank S. Campbell, H. Schneider, and a referee for supplying some important references.

REFERENCES

[1] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York-London 1969.

[2] D. Ž. ĐOKOVIĆ, *Extreme rays of certain cones of hermitian forms*, Proc. Amer. Math. Soc., 83 (1981), pp. 243–247.

[3] S. FRIEDLAND, *A characterization of transform absolute norms*, Linear Algebra and Appl., 28 (1979), pp. 63–68.

[4] W. GIVENS, *Fields of values of a matrix*, Proc. Amer. Math. Soc., 3 (1952), pp. 206–209.

[5] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York-Toronto-London, 1964.

[6] M. G. KREIN AND JU. L. ŠMUL'JAN, *Plus-operators in a space with indefinite metric*, Mat. Issled., 1(1) (1966), pp. 131–161; Amer. Math. Soc. Transl., (2) 85 (1969), pp. 93–113.

[7] R. LOEWY AND H. SCHNEIDER, *Positive operators on the n-dimensional ice-cream cone*, J. Math. Anal. Appl., 49 (1975), pp. 375–392.

[8] J. L. MOTT AND H. SCHNEIDER, *Matrix algebras and groups relatively bounded in norm*, Arch. Math., 10 (1959), pp. 1–6.

[9] B. D. SAUNDERS AND H. SCHNEIDER, *Norms and numerical ranges in finite dimensions*, unpublished.

[10] H. SCHNEIDER AND M. VIDYASAGAR, *Cross-positive matrices*, SIAM J. Numer. Anal., 7 (1970), pp. 508–519.

[11] J. S. VANDERGRAFT, *Spectral properties of matrices which have invariant cones*, SIAM J. Appl. Math., 16 (1968), pp. 1208–1222.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

005

DISCRETE TIME-BAND LIMITING OPERATORS AND COMMUTING TRIDIAGONAL MATRICES*

RONALD KEITH PERLINE†

Abstract. Time-band limiting operators, corresponding to classical discrete orthogonal families, admit commuting second order difference operators. A new proof is presented.

Key words. time-band limiting operators, commuting tridiagonal matrices

AMS(MOS) subject classification. 33A70

1. Introduction. Let $p_i(x)$ $\{i, x = 0, 1, 2, \cdots, N\}$ be a collection of linearly independent discrete functions orthogonal with respect to some weight $w(x)$. Following [G2], [P1] and [P2], we consider the analogue of "time limiting" and "band limiting" for ordinary Fourier analysis on the line. Let $f(x)$ be any discrete function; denote by M_K the operator which multiplies f by the characteristic function of the set $\{0, 1, \cdots, K\}$. We call M_K a "time limiting" operator. Similarly, denote by P_L the operator which projects f onto the span of the functions $\{p_0, p_1, \cdots, p_L\}$. P_L is a "band limiting" operator. The self-adjoint composition $M_KP_LM_K$ we call a "time-band limiting" operator.

These operators are analogues of the finite convolution operator $T:L^2 \to L^2$ given by

$$Tf(x) = \int_{-A}^{A} \frac{\sin B(x-y)}{x-y} f(y)\,dy,$$

which was subject to intensive study in the celebrated series of papers by Slepian, Landau and Pollak [S1], [S2], [S3], [S4], [S5]. The analysis of this operator was facilitated by a fortunate "accident": the existence of a commuting, self-adjoint, second order differential operator.

In the case of our (matrix) operator $M_KP_LM_K$, we could similarly hope for the existence of a commuting tridiagonal *matrix*. Perlstadt, motivated by the work of Grunbaum in [G1] and [G2], has shown that such a commuting tridiagonal matrix does indeed exist if the polynomial family $\{p_i(x)\}$ is of classical type: Poisson–Charlier, Meixner, Krawtchouk and Hahn (see [P1]). This result was generalized in [P2] to include the q-Racah polynomials of Askey and Wilson [A1], which include the classical families as special cases.

The proofs in [P1] and [P2] are strongly patterned after the proof given in [G2]. In particular, the properties that the q-Racah polynomials enjoy that appear in the proof are

(i) The p_i's are eigenfunctions of a second order difference operator;

(ii) Existence of a Christoffel–Darboux formula (equivalent to a three-term recursion relation);

(iii) Existence of a first order difference formula for p_i in terms of p_i and p_{i-1}.

The proof in [P2] is combinatorial and rather involved. This is due to the fact that the q-Racah polynomials appear in the calculations explicitly; and formulas involving the q-Racah polynomials tend toward the baroque, to say the least. It came as a pleasant surprise, therefore, when we discovered a simple, direct, constructive proof of the existence of a commuting tridiagonal matrix for the time-band limiting operator whenever the orthogonal family $\{p_i(x)\}$ satisfies properties (i) and (ii). At first glance, this result might

* Received by the editors March 3, 1986; accepted for publication June 4, 1986.

† Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104.

seem more general than that of [P2]; but according to [L1], any orthogonal family with these properties is in fact of q-Racah type.

2. The proof. First we establish some notation, and state our hypotheses in a manner that facilitates the proof of our theorem. As stated in the introduction, let $\{p_i(x)\}$ be our family of linearly independent discrete functions, orthogonal with respect to the weight $w(x)$. We assume that

(i) $D(p_i)(x) = \lambda(i)p_i(x)$; the functions $\{p_i\}$ are eigenfunctions for some second order difference operator D, self-adjoint with respect to the weight w;

(ii) There exists a discrete function $\Theta(x)$ that satisfies a three-recursion relation with respect to the functions $\{p_i(x)\}$:

$$\Theta(x)p_i(x)=a_ip_{i+1}(x)+b_ip_i(x)+c_ip_{i-1}(x).$$

Because we assume that the index i has the finite range $0 \leqq i \leqq N$, we impose the conditions that $a_N = 0$, $c_0 = 0$.

For example, the Hahn polynomials $h_i(x)$ (see [A2]) are defined by

$$h_i(x)={}_2F_1(-i,-x,i+\alpha+\beta+1;-N,\alpha+1;1),\qquad \alpha,\beta>1,\quad x,i=0,1,\cdots,N,$$

where ${}_2F_1$ is the generalized hypergeometric function. If we use $C(n, k)$ to denote the standard binomial coefficient, the weight associated with the family $\{h_i\}$ is

$$w(x)=\frac{C(\alpha+x,x)C(\beta+N-x,N-x)}{C(N+\alpha+\beta+1,N)}.$$

The Hahn polynomials satisfy

(i) Second order difference equation

$$(1/w(x))\Delta_+[-w(x-1)(\alpha+x)(N+1-x)\Delta_-h_i(x)]=(i)(i+\alpha+\beta+1)h_i(x),$$

where

$$\Delta_+f(x)=f(x+1)-f(x),\qquad \Delta_-f(x)=f(x)-f(x-1);$$

(ii) Three term recursion relation: if $p_i = h_i/\|h_i\|$, then the normalized p_i satisfy $xp_i(x) = a_ip_{i+1}(x) + b_ip_i(x) + c_ip_{i-1}(x)$; where $a_{i-1} = c_i$, and

$$a_i=\left(\frac{(i+1)(i+1+\beta)(i+\alpha+\beta+N+2)(i+\alpha+\beta+1)(i+\alpha+1)(N-i)}{(2i+\alpha+\beta+2)(2i+\alpha+\beta+3)(2i+\alpha+\beta+1)(2i+\alpha+\beta+2)}\right)^{1/2};$$

this is just a reformulation of the Christoffel–Darboux formula.

Returning to our main discussion: by replacing $p_i(x)$ by $p_i(x)\sqrt{w}(x)$ and D by $\sqrt{w}\,D\,1/\sqrt{w}$ (note that this is operator conjugation), we may assume without loss of generality that $w(x)\equiv 1$. It is also convenient to assume that the functions $\{p_i\}$ are normalized, so that they are orthonormal. With these assumptions, we note that $c_i = a_{i-1}$.

Introduce a second basis

$$d_i(x)=\begin{cases}1, & x=i,\\ 0 & \text{otherwise.}\end{cases}$$

Note that, for any function f, we have

$$f=\sum\langle f,d_i\rangle d_i=\sum f(i)d_i.$$

Conditions (i) and (ii) can be reformulated as follows:

(i) The linear transformation D is symmetric tridiagonal with respect to the d-basis $B_d = \{d_0, d_1, d_2, \cdots, d_N\}$; it is diagonal with respect to the p-basis $B_p = \{p_0, p_1, \cdots, p_N\}$.

(ii) The linear transformation $\Theta: f(x) \to \Theta(x)f(x)$ is diagonal with respect to B_d; it is symmetric tridiagonal with respect to B_p.

As in the introduction, let M_K denote projection onto the span of $\{d_0, d_1, \cdots, d_K\}$ and let P_L denote projection onto $\{p_0, p_1, \cdots, p_L\}$.

THEOREM. *There exists a symmetric tridiagonal matrix commuting with $M_K P_L M_K$.*

Proof. Consider the operator

$$\begin{array}{ccccc}(\Theta D + D\Theta) & - & (\Theta(K+1) + \Theta(K))D & - & (\lambda(L+1) + \lambda(L))\Theta = T(K, L) = T = \\ \text{(I)} & - & \text{(II)} & - & \text{(III)}.\end{array}$$

The matrix representation of T with respect to B_p is symmetric tridiagonal. In fact, its matrix representation looks like

$$\begin{array}{|ccccc:cccc|}
(0) & & & & (L) & (L+1) & & & \\
* & * & & & & & & & \\
* & * & * & & & & & & \\
 & * & * & * & & & & & \\
 & & * & * & * & & & & \\
 & & & * & * & 0 & & & \\
\hdashline
 & & & & 0 & * & * & & \\
 & & & & & * & * & * & \\
 & & & & & & * & * & * \\
 & & & & & & & * & *
\end{array}.$$

To see this, note that terms (I), (II), and (III) are all symmetric tridiagonal with respect to B_p, so T is also. In fact, (II) is diagonal with respect to B_p. Now consider the difference (I)–(III). It is easy to see that the $(L, L+1)$ entry of $(\Theta D + D\Theta)$ is just $(\lambda(L+1) + \lambda(L))$ times the $(L, L+1)$ entry of Θ; but this shows that the $(L, L+1)$ entry of (I)–(III) is zero, resulting in the matrix given above.

Similarly, the representation of T with respect to B_d looks like

$$\begin{array}{|ccccc:cccc|}
(0) & & & & (K) & (K+1) & & & \\
* & * & & & & & & & \\
* & * & * & & & & & & \\
 & * & * & * & & & & & \\
 & & * & * & * & & & & \\
 & & & * & * & 0 & & & \\
\hdashline
 & & & & 0 & * & * & & \\
 & & & & & * & * & * & \\
 & & & & & & * & * & * \\
 & & & & & & & * & *
\end{array}.$$

T commutes with P_L; this follows immediately from the matrix representation of T with respect to B_p given above, and the fact that the representation of P_L with respect to B_p is just

$$\begin{array}{|ccccc:ccc|}
(0) & & & & (L) & (L+1) & & \\
1 & & & & & & & \\
 & 1 & & & & & & \\
 & & 1 & & & & & \\
 & & & 1 & & & & \\
 & & & & 1 & & & \\
\hdashline
 & & & & & 0 & & \\
 & & & & & & 0 & \\
 & & & & & & & 0
\end{array}.$$

Similarly, T commutes with M_K, which can be seen by considering their representations with respect to B_d. Thus T commutes with $M_K P_L M_K$. The same type of argument shows that T is symmetric tridiagonal with respect to B_d and commutes with $P_L M_K P_L$. □

3. Remarks. One might ask how the commuting tridiagonal matrices generated here compare with those found in [P1] and [P2]. We report that a case-by-case check of the known circumstances under which commuting tridiagonal matrices (or the continuous analogues in [G2]) exist shows that our method produces the same commuting operator (F. A. Grunbaum and M. Reach, private communication).

Finally, since the motivation for finding the commuting tridiagonal matrix is to facilitate the spectral analysis of the operator $M_K P_L M_K$, it is desirable to have conditions which insure that the matrix T has simple spectrum. We state some simple sufficient conditions:

(i) The off-diagonal elements in the three-term recursion relation for Θ are nonzero;

(ii) The eigenvalues of the difference operator D satisfy

$$\lambda(L+1)+\lambda(L)=\lambda(K+1)+\lambda(K)\rightarrow L=K.$$

These conditions are satisfied for the classical orthogonal polynomials. We leave the proof that these conditions imply simple spectrum as an exercise for the interested reader.

REFERENCES

[A1] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, SIAM J. Math. Anal., 10 (1979), pp. 1008–1016.

[A2] R. Askey, *Orthogonal Polynomials and Special Functions*, CBMS Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[G1] F. A. Grunbaum, L. Longhi and M. Perlstadt, *Differential operators commuting with finite convolution operators: some nonabelian examples*, SIAM J. Appl. Math., 42 (1982), pp. 941–952.

[G2] F. A. Grunbaum, *A new property of reproducing kernels for classical orthogonal polynomials*, J. Math. Anal. Appl., 95 (1983), pp. 491–500.

[L1] D. Leonard, *Orthogonal polynomials, duality, and association schemes*, SIAM J. Math. Anal., 13 (1982), pp. 656–663.

[P1] M. Perlstadt, *Chopped orthogonal polynomial expansions—some discrete cases*, this Journal, 4 (1983), pp. 94–100.

[P2] ———, *A property of orthogonal polynomial familes with polynomial duals*, SIAM J. Math. Anal., 15 (1984), pp. 1043–1054.

[S1] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: I, Bell System Tech. J., 40 (1961), pp. 43–64.

[S2] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: II, Bell System Tech. J., 40 (1961), pp. 65–84.

[S3] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: III, Bell System Tech. J., 41 (1962), pp. 1295–1336.

[S4] D. Slepian, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: IV, Bell System Tech. J., 43 (1964), pp. 3009–3058.

[S5] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: V, Bell System Tech. J., 57 (1978), pp. 1371–1430.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

006

AN EXTENSION OF PLACKETT'S DIFFERENTIAL EQUATION FOR THE MULTIVARIATE NORMAL DENSITY*

SIMEON M. BERMAN†

Abstract. Let $f(\mathbf{x}, \mathbf{y}; B)$, with $\mathbf{x}$, $\mathbf{y}$ in R^m, and B a nonsingular real $m \times m$ matrix, be a function of the form

$$f = (2\pi)^{-m/2}|\det B|^{-1/2} \exp(-\tfrac{1}{2}\mathbf{x}'B^{-1}\mathbf{y}).$$

It is shown that f satisfies a partial differential equation which represents a generalization of Plackett's equation in the case where B is positive definite, that is, where f is a normal density in m dimensions.

Key words. nonsingular matrix, positive definite matrix, multivariate normal density, partial differential equation, Gaussian process

AMS(MOS) subject classifications. 62H05, 60E99, 15A24, 60G15

1. Introduction and summary. Let $f(x_1, \cdots, x_m; (b_{ij}))$ be the m-variable normal density function with mean vector $\mathbf{0}$ and positive definite covariance matrix (b_{ij}). Extending a known result in the case $m = 2$, Plackett [5] discovered the general relation

$$\frac{\partial f}{\partial b_{ij}} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \qquad i \neq j. \tag{1.1}$$

His proof is based on the inversion formula for the characteristic function

$$f(x_1, \cdots, x_m; (b_{ij})) = (2\pi)^{-m} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(i \sum_{j=1}^{m} x_j z_j - \frac{1}{2} \sum_{j,k=1}^{m} b_{jk} z_j z_k\right) \prod_j dz_j. \tag{1.2}$$

The result (1.1) follows by taking the appropriate derivatives under the integral in (1.2). While the differentiation with respect to x_i is justified by basic properties of the integral, I found that the argument for differentiation with respect to b_{ij} required much more. Indeed, for $h > 0$, consider the difference quotient of increment h leading to $\partial f/\partial b_{ij}$. It is obtained from the integral in (1.2) by multiplying the integrand by the factor $h^{-1}(1 - \exp(-\frac{1}{2}hz_i z_j))$. There is no obvious dominating function for this factor in the region where $z_i z_j < 0$, and so the dominated convergence theorem cannot be applied without more delicate estimates of the remaining part of the integrand. These require some of the deeper properties of positive definite forms.

A complete proof of (1.1) has apparently never been published. In view of its importance in the theory of extremes of stationary Gaussian processes (see Berman [1], Galambos [2], Leadbetter, Lindgren and Rootzen [4]), such a proof should be available in the literature. The inequality which has become known as "Slepian's inequality" [6] is also based on (1.1).

It is the purpose of this note to present a simple algebraic proof of a more general version of (1.1). In the place of the multivariate normal density f, we consider the more general function

$$f(x_1, \cdots, x_m, y_1, \cdots, y_m; (b_{ij})) = (2\pi)^{-m/2}|\det B|^{-1/2} \exp\left(-\frac{1}{2} \sum_{i,j=1}^{m} a_{ij} x_i y_j\right), \tag{1.3}$$

where $B = (b_{ij})$ is nonsingular with inverse $A = (a_{ij})$. B is not necessarily positive definite or even symmetric. Our result is that f satisfies the system of equations

* Received by the editors January 27, 1986; accepted for publication July 23, 1986. This paper represents results obtained at the Courant Institute of Mathematical Sciences, New York University, under the sponsorship of National Science Foundation grant DMS 85 01512.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012.

$$(1.4)\qquad \frac{\partial f}{\partial b_{hk}}=\frac{\partial^2 f}{\partial x_k \partial y_h}+\frac{1}{f}\cdot\frac{\partial f}{\partial x_k}\cdot\frac{\partial f}{\partial y_h},\qquad h,k=1,\cdots,m.$$

(The derivative is assumed to be defined in the domain of the variables b_{ij} where the matrix is nonsingular.) In the particular case where $x_k = y_k$ for all k, and B is symmetric (not necessarily positive definite), (1.4) is replaced by the original relation (1.1).

In a recent paper Joag-Dev, Perlman and Pitt [3], assuming the validity of (1.1), obtained some extensions of it. It can be shown that (1.1) itself can actually be derived from their formulas (6) and (7). Indeed, we will show that the more general relation (1.4) can be obtained by a simple extension of the elementary algebraic methods which they used. As a consequence, this finally furnishes an explicit proof of (1.1).

2. Proof of (1.4). We employ the more general versions of the relations in [3], formula (6), for nonsingular but not necessarily symmetric matrices. From the relation

$$\frac{\partial}{\partial b_{hk}}(\det B)=(\det B)a_{kh},$$

we obtain

$$\frac{\partial}{\partial b_{hk}}|\det B|=|\det B|a_{kh},$$

and, from the latter,

$$(2.1)\qquad \frac{\partial}{\partial b_{hk}}|\det B|^{-1/2}=-\frac{1}{2}|\det B|^{-1/2}a_{kh}.$$

Next we show that

$$(2.2)\qquad \frac{\partial}{\partial b_{hk}}\sum_{i,j}a_{ij}x_iy_j=-\sum_i a_{ih}x_i\cdot\sum_j a_{kj}y_j.$$

For the proof of this elementary result in the general nonsymmetric case, write $B=(b_{ij}(t))$, where $b_{ij}(t)$ is a differentiable function of t, and $\partial B/\partial t=(\partial b_{ij}(t))$. Differentiate both members of the equation $BA = I$ with respect to t, and then multiply the resulting equation by A to obtain $\partial A/\partial t = -A(\partial B/\partial t)A$. In the particular case where the variable t is b_{hk}, $\partial B/\partial b_{hk}$ is the matrix with entry 1 in position (h, k) and 0 elsewhere, and so $\partial a_{ij}/\partial b_{hk} = -a_{ih}a_{kj}$, and (2.2) follows. Finally we have

$$(2.3)\qquad \frac{\partial}{\partial x_k}\sum_{i,j}a_{ij}x_iy_j=\sum_j a_{kj}y_j,\qquad \frac{\partial}{\partial y_h}\sum_{i,j}a_{ij}x_iy_j=\sum_i a_{ih}x_i.$$

The result (1.4) now follows from the form of f in (1.3) through elementary calculations using (2.1), (2.2) and (2.3).

REFERENCES

[1] S. M. BERMAN, *Limit theorems for the maximum term in stationary sequences*, Ann. Math. Statist., 35 (1964), pp. 502–516.

[2] J. GALAMBOS, *The Asymptotic Theory of Extreme Order Statistics*, John Wiley, New York, 1978.

[3] K. JOAG-DEV, M. PERLMAN AND L. PITT, *Association of normal random variables and Slepian's inequality*, Ann. Probab., 11 (1983), pp. 451–455.

[4] M. R. LEADBETTER, G. LINDGREN AND H. ROOTZEN, *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York, 1983.

[5] R. L. PLACKETT, *A reduction formula for multivariate normal integrals*, Biometrika, 41 (1954), pp. 351–360.

[6] D. SLEPIAN, *The one-sided barrier problem for Gaussian noise*, Bell System Tech. J., 41 (1962), pp. 463–501.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

007

EMBEDDING OUTERPLANAR GRAPHS IN SMALL BOOKS*

LENWOOD S. HEATH†

Abstract. We investigate the problem of embedding graphs in books. A *book* is some number of half-planes (the *pages* of the book), which share a common line as boundary (the *spine* of the book). A *book embedding* of a graph embeds the vertices on the spine in some order and embeds each edge in some page so that in each page no two edges intersect. The *pagenumber* of a graph is the number of pages in a minimum-page embedding of the graph. The *pagewidth* of a book embedding is the maximum cutwidth of the embedding in any page. A practical application of book embeddings is in the realization of a fault-tolerant array of VLSI processors.

Our result is an $O(n \log n)$ time algorithm for embedding an n-vertex outerplanar graph with small pagewidth. The algorithm embeds any d-valent outerplanar graph in a two-page book with $O(d \log n)$ pagewidth. This result is optimal in pagewidth to within a constant factor for the class of outerplanar graphs. As there are trivalent outerplanar graphs that require $\Omega(n)$ pagewidth in any one-page embedding, the pagenumber of our embedding is exactly optimal for the stated pagewidth. The significance for VLSI design is that any outerplanar graph can be implemented in small area in a fault-tolerant fashion.

Key words. outerplanar graphs, book embedding, algorithm, hamiltonian cycles

AMS(MOS) subject classifications. 05C45, 68Q35, 94C15

1. The problem. We study embeddings of graphs in structures called *books*. A *book* consists of a *spine* and some number of *pages*. The spine of a book is a line. For simple exposition, view the spine as being horizontal. Each page of the book is a half-plane that has the spine as its boundary. Thus any half-plane is a one-page book, and a plane with a distinguished horizontal line is a two-page book.

The embedding of an undirected graph in a book consists of two steps. The first step places the vertices of the graph on the spine in some order. The second step assigns each edge of the graph to one page of the book in such a way that on each page, the edges assigned to that page *do not cross*. Whether two edges cross is determined by the order of the vertices. If (s, t) and (u, v) are edges of the graph with $s < u < v$ and $s < t$, then the edges cross if and only if $s < u < t < v$. The resulting embedding is called a *book embedding* of the graph.

There are two measures of the quality of a book embedding for G.

The first measure is the *pagenumber* of the embedding, which is the number of pages in the book.

The *pagenumber* of the graph G is the minimum pagenumber of any book embedding of G. The *pagenumber* of a class of graphs is the minimum number of pages that every member of the class can be embedded in, as a function of graph size. The *width* of a page is the maximum number of edges that intersect any half-line perpendicular to the spine in the page.

The second measure is the *pagewidth* of the embedding, which is the maximum width of any page.

The *pagewidth* of the graph G is the minimum pagewidth of any book embedding of G in a book having a minimum number of pages. The *pagewidth* of a class of graphs is the minimum pagewidth that every member of the class can be embedded in, as a function of graph size. The *book embedding problem* is to find good book embeddings for a graph family with respect to one or both of these measures.

* Received by the editors November 13, 1985; accepted for publication (in revised form) July 31, 1986. This research was supported by National Science Foundation grant MCS-83-01213.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

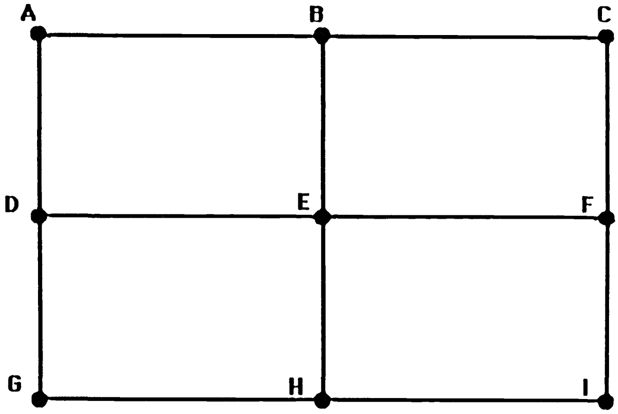


FIG. 1. *Grid graph G.*

As an example, consider the grid graph G of Fig. 1. A two-page embedding of G is shown in Fig. 2. The vertices of G are placed on the spine in the order A-B-C-F-E-D-G-H-I. The first page consists of the upper half-plane, and the second page consists of the lower half-plane. Edge (B, E) of the first page crosses edge (F, I) of the second page, so these two edges cannot be assigned to the same page of this book. The pagenumber of the book embedding is two, and the pagewidth is three as witnessed by the nested edges (A, D), (B, E), and (C, F) (both measures are optimal for G).

The book embedding problem is of interest because it models problems in several areas of computer science and VLSI theory. We mention here only problems arising from the DIOGENES approach of Rosenberg [9]. For further motivations, see Heath [5] or Chung, Leighton and Rosenberg [2].

Rosenberg [9] proposes the DIOGENES approach to the design of fault-tolerant arrays of VLSI processors. The elements of the approach are sketched here. One lays out some number of identical processors in a (conceptual) line. One provides sufficiently many processors so that one expects (probabilistically) that enough good processors exist to implement the desired array.

Bundles of wires with embedded switches run parallel to the line of processors. Each bundle is capable of implementing a hardware stack of connections among processors. Each connection occurs on exactly one hardware stack (bundle). For any processor, a connection to a processor on its right is pushed on a stack; each connection to a processor on its left is popped from a stack. In this way, each connection to a good processor requires one stack operation at that processor. No stack operations occur at a bad processor. Since the state of a processor as good or bad is a binary value, a single control signal can cause the shift (push or pop) of many connections. Thus, fault tolerance is achieved by switching in only good processors.

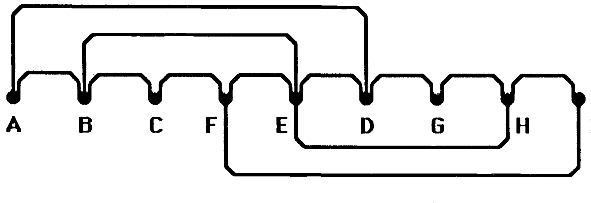


FIG. 2. *Two-page embedding of G.*

The desired array of processors is modeled as a *connection graph*; the vertices represent the processors, and the edges represent the desired connections between processors. The DIOGENES design problem is to determine the number of stacks and the *stackwidths* (the number of connections carried by each stack) required to implement the array of processors. In a way analogous to a hardware stack, it is possible to view one page of a book embedding as a stack of edges. For any vertex, each incident edge that connects it to a vertex to its right is pushed on a stack; each incident edge that connects it to a vertex to its left is popped from a stack. The DIOGENES design problem for an array of processors is exactly the book embedding problem for the corresponding connection graph. The number of stacks is exactly the number of pages. The stackwidths are the widths of the pages.

In this paper, we consider the problem of simultaneously attaining small pagenumber and small pagewidth. We consider the class of *outerplanar* graphs (an outerplanar graph is one that has a planar embedding with all vertices on the exterior face). There exist outerplanar graphs of size n that have pagewidth $\Omega(n)$ in any one-page embedding but have pagewidth $O(1)$ in an optimal two-page embedding. These graphs exhibit a tradeoff between pagenumber and pagewidth. We present an algorithm that produces a two-page embedding of small pagewidth for any outerplanar graph. The pagewidth that is attainable depends partly on the *valence* (maximum degree) of the outerplanar graph. Let G be a d-valent outerplanar graph with n vertices. Our algorithm embeds G in a two-page book having pagewidth less than $Cd \log n$ where $C = 8/(\log 3/2)$ (all logarithms are to the base two). This result is within a constant factor of optimal in pagewidth for the class of outerplanar graphs. The algorithm executes in time $O(n \log n)$.

The remainder of the paper consists of seven sections. In the next section, we survey previous results on book embeddings relevant to our algorithm. In § 3, we discuss tradeoffs between pagenumber and pagewidth and give an example of such a tradeoff. Section 4 presents the essential ideas of the algorithm, while § 5 gives the detailed statement. Section 6 proves the correctness of the algorithm, and § 7 establishes its performance. In the last section, we conclude with a discussion of the significance of our result and suggest an area for further research.

2. Previous results. The original statement of book embedding is a linear embedding performed in two parts. First, the vertices of a graph are placed on a line in some order. Second, each edge of the graph is embedded in one page so that no edges in the same page cross.

The resulting linear embedding can be transformed into a circular embedding in three steps. First, choose a distinct color for each page of the book, and assign each edge the color of its page. Second, "close" the book by projecting all pages (and their edges) into a single page. In this one-page book, if two edges cross, then the two edges have different colors. Third, curve the spine into a circle so that the "ends" at infinity are identified.

The result of the transformation is an alternate two-part formulation of the book embedding problem. First, order the vertices of the graph on a circle. Second, draw the edges of the graph as chords of the circle. Color the chords (edges) so that if two chords intersect in the interior of the circle, the chords have different colors. The number of colors in the circular embedding is exactly the number of pages in the corresponding linear embedding.

A useful consequence of the circular formulation is that any p-page graph is a subgraph of a p-page hamiltonian graph. (A graph is *hamiltonian* if it has a cycle containing all its vertices; such a cycle is called a *hamiltonian cycle.*) Moreover, the order of the vertices in the circular embedding is exactly the order of the vertices in the hamiltonian cycle. To see this, let $v_1, v_2, \cdots, v_n$ be the vertices of the p-page graph in the cyclic order of the circular embedding. Add each of the edges (chords) (v_k, v_{k-1}), $1 \leqq k \leqq n$ (where $k-1$ is taken modulo n) that are not already present. Since these edges connect vertices adjacent on the circle, they cannot intersect any other edges. Therefore, each of the edges can legitimately be assigned to any page. The resulting edge-augmented graph is a p-page graph, with hamiltonian cycle $v_1, \cdots, v_n$.

The idea of adding edges to a graph to obtain a hamiltonian cycle is a strategy for obtaining the vertex order of a book embedding. We will call a cycle obtained in this fashion *superhamiltonian*. The following heuristic for book embedding a graph G is proposed in [2]:

(1) Obtain a superhamiltonian cycle for G and place the vertices of G on the circle in the order of the cycle;

(2) Color the edges of G by coloring the associated circle graph.

Finding an optimal solution to the second step in the heuristic is an NP-complete problem (Garey et al. [3]). The first step can be done in a number of ways; in fact, any ordering of the vertices can be obtained for a superhamiltonian cycle by adding the right edges. Thus, the problem of finding good book embeddings can be approached as that of finding a superhamiltonian cycle in an intelligent fashion so that a good (but not necessarily optimal) edge coloring can be produced.

2.1. One-page graphs. Any one-page graph can be embedded in the plane so that its vertices are on the spine and its edges are in the first page (the upper half-plane). Then all its vertices are exposed to the lower half-plane, which is a subset of the exterior face of the embedding. Thus the graph is outerplanar.

One characterization of an outerplanar graph is that its vertices can be embedded on a circle so that all its edges are inside the circle and no two edges intersect. This is just the condition that the graph be one-page embeddable under the circular formulation. We have the following:

PROPOSITION 1 (Bernhart and Kainen [1]). *G is one-page embeddable if and only if it is outerplanar.*

In fact, a k-page embedding of a graph G yields a decomposition of G into k outerplanar subgraphs, one for each page. The subgraphs share the vertices of G but are edge-disjoint. The outerplanarity of each subgraph is witnessed by the same circular ordering as that of the original book embedding.

2.2. Two-page graphs. Each two-page graph is a subgraph of a two-page hamiltonian graph. Every two-page graph is planar since the two half-planes (pages) together form a plane. Thus a two-page graph is a subgraph of a planar Hamiltonian graph.

Define a graph to be *subhamiltonian* if it is the subgraph of a planar hamiltonian graph. Given a subhamiltonian graph G, it is easy to show that G has a two-page embedding ([1]). Edge-augment G to obtain a superhamiltonian cycle in a planar graph. Order the vertices of G on a circle according to the superhamiltonian cycle. The edges of G interior to the cycle form an outerplanar graph. The edges exterior to the cycle form another outerplanar graph with its vertices in the same order as those of the interior one. A two-page embedding of G results. Thus we have the following:

PROPOSITION 2 [1]. *G is two-page embeddable if and only if it is subhamiltonian.*

Propositions 1 and 2 are the results we use in our algorithm to obtain two-page embeddings of outerplanar graphs with small pagewidth. From Proposition 1, an outerplanar graph G has a one-page embedding with all edges embedded in the upper half-plane (page). Our algorithm adds edges to G in the lower half-plane so that a planar,

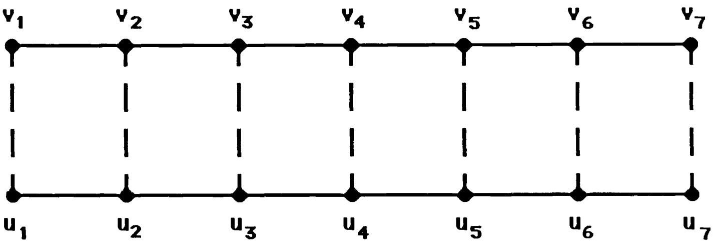


FIG. 3. *The 7-ladder L_7.*

hamiltonian supergraph results. By Proposition 2, the superhamiltonian cycle yields a two-page embedding of G with the vertices of G in cycle order.

3. Tradeoffs. We investigate the problem of tradeoffs between pagenumber and pagewidth in book embeddings. Motivation is best provided by an example from Chung, Leighton and Rosenberg [2]. The example is a sequence of outerplanar graphs $\{L_m\}$ for which any one-page embedding requires large pagewidth $\lceil m/2 \rceil$, but for which there exist two-page embeddings with pagewidth 2. The sequence consists of *m-ladders* (in [2], an m-ladder is called a *depth-m K_2-cylinder*). The m-ladder L_m has vertex set

$$\{u_1, \cdots, u_m\} \cup \{v_1, \cdots, v_m\}$$

and edge set

$$\{(u_k, u_{k+1}) | 1 \leqq k < m\} \cup \{(v_k, v_{k+1}) | 1 \leqq k < m\} \cup \{(u_k, v_k) | 1 \leqq k \leqq m\}.$$

The first two components of the edge set constitute the two *sides* of the ladder while the last component constitutes its *rungs*. Figure 3 illustrates L_7. The sides are solid and the rungs are dashed.

The m-ladder is clearly outerplanar and biconnected. By biconnectivity, L_m has a unique outerplanar embedding (Syslo [10]). Therefore, L_m has a unique one-page embedding up to reflection and circular permutation. Figure 4 illustrates a one-page embedding of L_7 of minimal pagewidth over all one-page embeddings. The rungs $\{(u_4, v_4), (u_5, v_5), (u_6, v_6), (u_7, v_7)\}$ nest over the interval (u_7, v_7). Hence the pagewidth is $\geqq 4$. A moment's reflection generalizes this observation: In any one-page embedding for L_m, at least $\lceil m/2 \rceil$ rungs nest over some interval; hence pagewidth is $\geqq \lceil m/2 \rceil$.

Figure 5 illustrates a two-page embedding for L_7 that has pagewidth 2. The corresponding superhamiltonian cycle is illustrated in Fig. 6. This superhamiltonian cycle is easily generalized, giving a two-page embedding of any L_m with pagewidth 2.

We now discuss tradeoffs in the general setting of an arbitrary graph G. Let P be the pagenumber of G. For each $p \geqq P$, there exist one or more embeddings of G in a p-

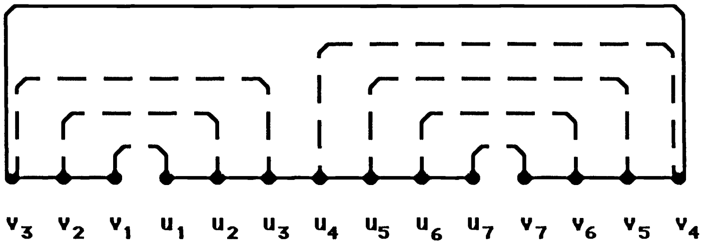


FIG. 4. *One-page embedding for L_7.*

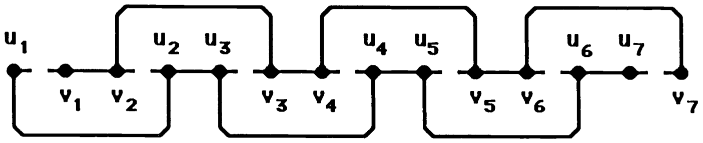


FIG. 5. *Two-page embedding for L_7.*

page book. Among all those p-page embeddings of G, let w_p denote the pagewidth of one having minimum pagewidth. These pagewidths are nonincreasing:

$$w_P \geqq w_{P+1} \geqq \cdots \geqq w_p, p \geqq P.$$

In the extreme case that $p \geqq |E|$, $w_p = 1$, as each edge may be assigned to a distinct page.

We are particularly interested in the product pw_p. We seek cases where pw_p is within a constant factor of the cutwidth of G. Note that pw_p is an upper bound on the cutwidth of the best p-page embedding of G. In the context of the DIOGENES approach, pw_p is an upper bound on the height of a p-stack DIOGENES layout of G. Hence, we seek DIOGENES layouts of G that are within a constant factor of optimal in area over all linear layouts and within a small additive constant of optimal in stacknumber.

Our result is for the class of one-page (i.e., outerplanar) graphs. The m-ladder exhibits an extreme pagewidth tradeoff between one-page and two-page embeddings. For general outerplanar graphs, we do not expect such an extreme tradeoff. Since there exist outerplanar graphs that have one-page embeddings of minimal pagewidth, e.g., complete binary trees, the tradeoff in going from one page to two pages can be arbitrarily small, even zero.

An n-vertex complete $(d-1)$-ary tree has cutwidth $\geqq(d/2)\log n$ (Lengauer [6]). (All logarithms are to the base 2.) Hence, any book embedding of a complete $(d-1)$-ary tree in a constant number of pages requires pagewidth $\Omega(d \log n)$. In general, we cannot assume that outerplanar graphs have pagewidth $o(\log n)$.

4. Overview of the algorithm. The tradeoff result we show is that any d-valent outerplanar graph G can be embedded in a two-page book with pagewidth $Cd \log n$, where $C = 8/(\log 3/2)$. From the observations in the preceding section regarding m-ladders and complete $(d-1)$-ary trees, this result is optimal in pagenumber and within a constant factor of optimal in pagewidth for the class of d-valent outerplanar graphs. We prove our result via a recursive algorithm.

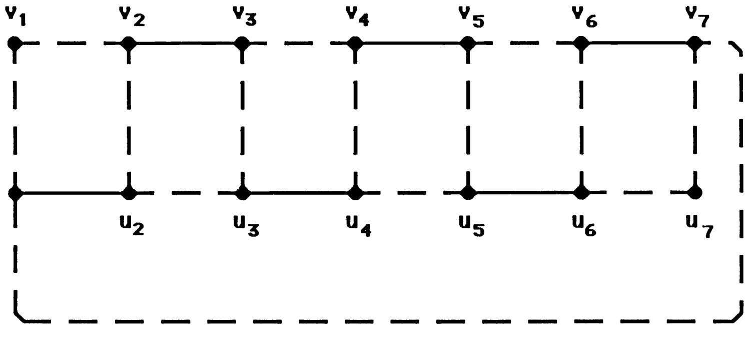


FIG. 6. *Superhamiltonian cycle for L_7.*

We aim for an algorithm that, when given an n-vertex d-valent outerplanar graph, returns a two-page embedding with pagewidth logarithmic in n. The input and output requirements of such an algorithm are a useful place to start.

The input to the algorithm is a d-valent outerplanar graph $G = (V, E)$. The manner of representing this input should witness the outerplanarity of G. Hence, a one-page embedding of G is the required form for the input. The linearization of V orders the vertices and provides names $1, 2, \cdots, n$ for the vertices. The order of the vertices in the two-page embedding will *not* be the original order, but we shall continue to use the *names*. Since the algorithm is recursive, the same vertex will have different names at different levels of recursion. Figure 7 illustrates a possible form of the input when $G = L_7$.

The output of the algorithm is a two-page embedding of G with logarithmic pagewidth. To give a two-page embedding for G, it is sufficient to give a superhamiltonian cycle H in a supergraph G' of G (Proposition 2). $G' = (V, E)$ is actually a *multigraph* (i.e., it may contain loops and multiple copies of edges) that contains all the edges of G plus possibly edges added to obtain H. $H \subset E$ is a set of n edges; since H is superhamiltonian, each of $1, \cdots, n$ appears exactly twice among these edges. H represents $2n$ different book embeddings for G: there are n choices for the leftmost vertex, and there are two directions to the cycle. The algorithm fixes the desired book embedding by returning the leftmost (x) and rightmost (y) vertices of the two-page embedding. We call x and y the *vertices of attachment* for G', for reasons that will become clear. The output of the algorithm is then the ordered triple $(G', H, (x, y))$.

We imagine the one-page embedding of G as follows. The vertices are on a horizontal line in a plane, and the edges are drawn in the upper half-plane. In general, there are many sets of edges that can be added to G without destroying planarity. We restrict ourselves to two types of edges, *upper edges* and *lower edges*, depending on which half-plane the edges are embedded in. (Thus our restriction is that no edge uses both half-planes in its embedding.) The original edges of G are always upper edges. The algorithm may add an upper edge if it will not cross an existing upper edge. The algorithm may add a lower edge if it will not cross an existing lower edge. In particular, we may (and shall) assume that the upper edges $(i, i + 1)$, $1 \leqq i < n$ are always present in G; they can always be added with no affect on pagenumber and at most unit increase in pagewidth.

The algorithm uses the divide-and-conquer paradigm. It determines subgraphs of G to work on separately before the results are joined together to obtain G'. Each subgraph is induced by a subinterval of $[1, n]$. We define the closed subinterval $[i, j]$ to be the set of integers $\{i, i + 1, \cdots, j\}$. We define two types of half-closed, half-open subintervals: $[i, j)$ denotes $[i, j - 1]$ and $(i, j]$ denotes $[i + 1, j]$. For any subinterval α, size $\{\alpha\}$ denotes the number of vertices in the subinterval; hence, size $\{[i, j]\} = j - i + 1$. Define $G[i, j]$

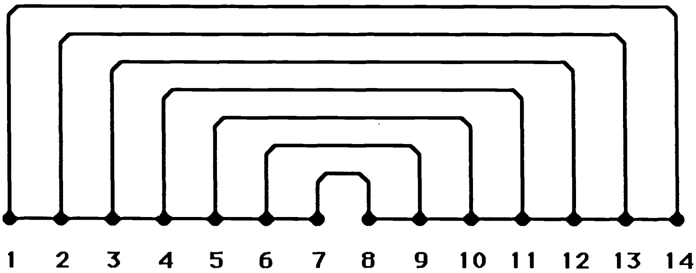


FIG. 7. *Input representation for L_7.*

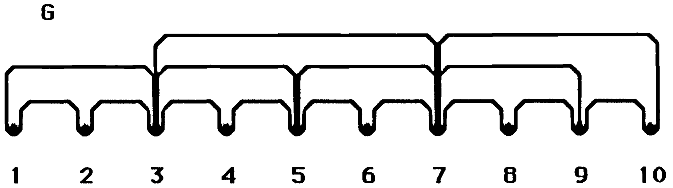


FIG. 8. *Sample G for divide-and-conquer.*

to be the subgraph of G induced on the vertices in the interval $[i, j]$. If the algorithm is applied to $G[i, j]$ the result is $(G'[i, j], H, (x, y))$, where (x, y) determines the first and last vertices of a two-page embedding of $G[i, j]$ with pagewidth logarithmic in size $\{[i, j]\}$.

The choice of subintervals depends on the structure of the one-page embedding of G. Define an *exposed vertex* w of G to be one for which G contains no (upper) edge (u, v) satisfying $u < w < v$. Thus an exposed vertex w is one that is "visible" from the infinite region of the upper half-plane. Each exposed vertex of G except 1 and n is a cutpoint of G whose removal separates G into left and right subgraphs.

An example will illustrate the divide-and-conquer paradigm. Figure 8 shows a sample G in a one-page embedding. The exposed vertices of G are 1, 3, 7 and 10. The algorithm recognizes that each of the edges (1, 3), (3, 7) and (7, 10) is "highest" in the sense that no other edge passes over it. These three edges determine three *nondisjoint* subintervals [1, 3], [3, 7] and [7, 10]. In order to decompose the interval into *disjoint* subintervals, the algorithm chooses the largest, [3, 7], to remain intact, and removes one vertex from each of the other two subintervals. The resulting subintervals are [1, 2], [3, 7] and [8, 10]. The algorithm recursively applies itself to each of the subintervals. The result to this point is shown in Fig. 9. Each subproblem displays a superhamiltonian cycle of its subgraph and the first and last vertices of the corresponding two-page embedding. In Fig. 10, these three superhamiltonian cycles are replaced by a superhamiltonian cycle for the entire graph. Lower edges (1, 2), (4, 6) and (8, 10) are deleted and lower edges (2, 4), (6, 8) and (1, 10) are added.

If two exposed vertices i and j are joined by an (upper) edge (i, j), then there are no other exposed vertices in the interval $[i, j]$. In this case, we call $G[i, j]$ a *block*, denoted $B[i, j]$. When the interval $[1, n]$ is partitioned into subintervals, there will be edges with endpoints in different subintervals. Such *dangling* edges are exactly those edges of G not in any of the subgraphs generated by the subintervals. In the case of a block $B[i, j]$, these dangling edges can be incident to only i or j. The total number of such edges incident to i or j is called the *edge deficit* of $B[i, j]$, denoted def $\{[i, j]\}$. It is always true that

$$\text{def}\,\{[i, j]\} \leqq 2(d-1).$$

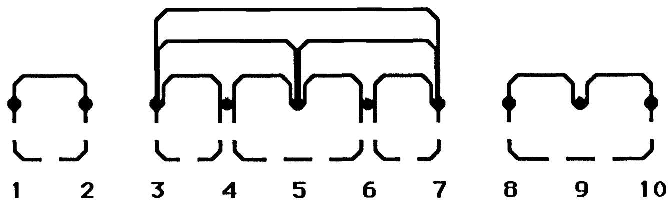


FIG. 9. *Results of subproblems.*

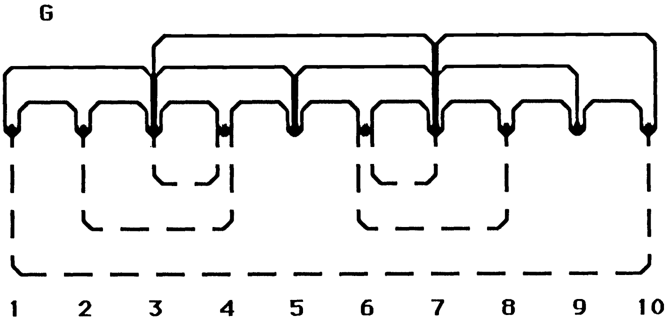


FIG. 10. *Superhamiltonian cycle for G.*

In Fig. 8, $B[1, 3]$, $B[3, 7]$ and $B[7, 10]$ are the blocks of G, and def $\{[3, 7]\} = 5$ (because of edges (1, 3), (2, 3), (7, 8), (7, 9), and (7, 10)).

The example of Figs. 8–10 illustrate the execution of the algorithm in the case that G has two or more blocks. There is another possible case: G has only one block. In that case, the divide-and-conquer construction is more complex. The two divide-and-conquer constructions corresponding to these two cases are developed in turn in the next two subsections.

4.1. String construction. We now describe one of the two constructions used to obtain a superhamiltonian cycle for G from superhamiltonian cycles for the graphs induced by subintervals. It is called the *string construction.* (The name suggests that the superhamiltonian cycles for the subintervals are *strung together* to obtain a superhamiltonian cycle for the entire interval.) It is employed when the number of exposed vertices is greater than two, i.e., when G is not one block. The partition into subintervals keys on the largest block, say $B[i, j]$: $B[i, j]$, is taken to be one of the subintervals.

A precise description of the partition into subintervals requires more notation. Let $m_1, m_2, \cdots, m_q$ be the exposed vertices of G in ascending order. Suppose $B[m_k, m_{k+1}]$ is the largest block in G. Figure 11 illustrates the situation. The partition into $q - 1$ subintervals is

$$\{[m_1, m_2), [m_2, m_3), \cdots, [m_{k-1}, m_k), [m_k, m_{k+1}], (m_{k+1}, m_{k+2}], \cdots, (m_{q-1}, m_q]\}.$$

Note that $B[m_k, m_{k+1}]$ is the only block of G in the partition. It is called the *key* block of the partition. The other subintervals are called *side* subintervals. Figure 12 illustrates the partition of G.

The algorithm is recursively applied to the jth subinterval to obtain

$$(G_j', H_j, (x_j, y_j)).$$

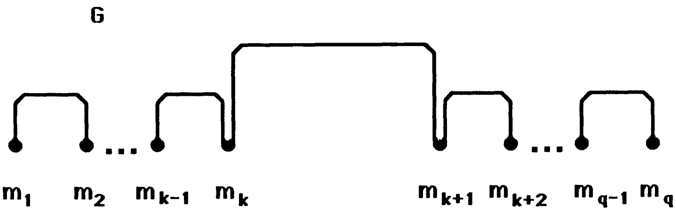


FIG. 11. *Exposed vertices.*

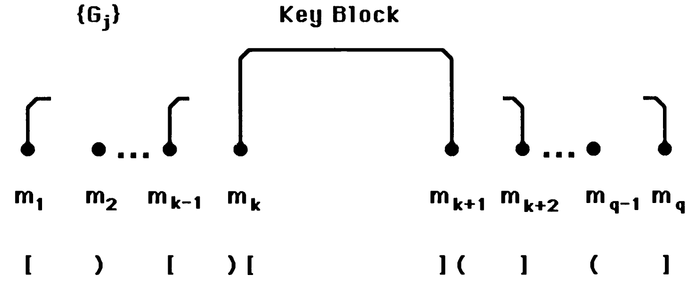


FIG. 12. *Partition into subintervals.*

G' is obtained in two steps. First, all edges added to the G_j' are added to G (Fig. 13). Second, the lower edges $\{(x_j, y_j)\}$ are deleted and the lower edges

$$\{(y_j, x_{j+1}) | 1 \leqq j \leqq q-2\} \cup \{(x_1, y_{q-1})\}$$

are added. H is obtained from $\cup H_j$ by deleting and adding the same edges (Fig. 14). Assigning $(x, y) = (x_1, y_{q-1})$ completes the string construction. The correctness of the construction is proved in Lemma 5 (§ 6).

4.2. Ladder construction. In this section, we consider the case when G has only one block, so we are unable to divide G into subintervals based on blocks. To reach a solution, we first focus on the problem of obtaining logarithmic pagewidth. To obtain logarithmic pagewidth, it is clearly sufficient that the linear layout corresponding to the two-page embedding have logarithmic cutwidth. An approach to small cutwidth is the recursive application of a *separator theorem* (see Lipton and Tarjan [7]). A separator theorem states that the removal of some number of vertices from a graph will partition the remainder of the graph into two subgraphs of approximately equal size. For outerplanar graphs, a two-vertex separator always exists.

LEMMA 3. *Let G be an outerplanar graph containing at least* 3 *vertices. There exist vertices x and y whose removal separates G into disjoint subgraphs G_1 and G_2 such that $\frac{1}{3}n < |G_k| < \frac{2}{3}n$, $k = 1, 2$. If (x, y) is not an upper edge of G, then it can be added to G as an upper edge without inducing a crossing.*

Proof. Since G is outerplanar, we can use the circular formulation of book embedding to embed G in a circle. The vertices of G are placed equally spaced on the circle. The edges of G are chords of the circle with no two chords intersecting. If the center of the circle lies on an edge, let x and y be the endpoints of the edge; in this case, $|G_k| < \frac{1}{2}n$, $k = 1, 2$, and the result follows. Otherwise, let F be the face of G containing the center. If two vertices on F are on a diameter, let them be x and y, and the result follows.

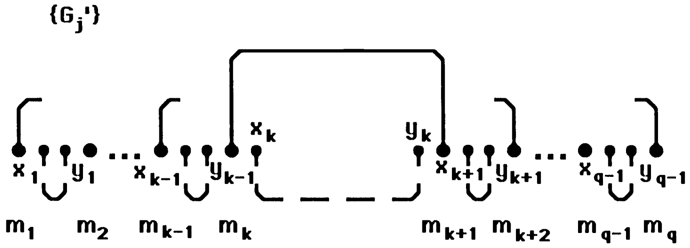


FIG. 13. *Results for subintervals.*

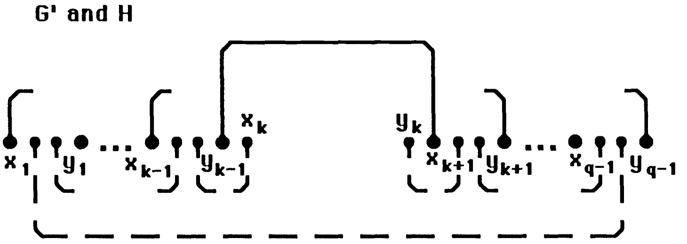


FIG. 14. *Subintervals strung together.*

Otherwise, triangulate F within the circle. The center of the circle lies within some resulting triangle (u, v, w). Let the angle $\angle uvw$ be the largest of the triangle. This angle is easily seen to be between 60° and 90°. Let $x = u$ and $y = w$. Let G_1 be the graph induced by the vertices within the angle $\angle uvw$, and let G_2 be the graph induced by the vertices outside the angle $\angle uvw$. Then the removal of x, y separates G into G_1 and G_2 where

$$\tfrac{1}{3}n < |G_1| < \tfrac{1}{2}n.$$

Note that the edge (x, y) can be added to G without destroying the outerplanar embedding. The lemma follows. □

If (x, y) is not already an edge of G, it can be added without destroying outerplanarity. An edge (x, y) that satisfies Lemma 3 is called a *separating edge.* An algorithm to obtain logarithmic cutwidth for a d-valent outerplanar graph G can select a separating edge (x, y) and apply itself recursively to the resulting G_1 and G_2. However, it is unclear how to obtain a superhamiltonian cycle for G from superhamiltonian cycles for G_1 and G_2.

Our algorithm uses separating edges in another way so as to make it possible to derive a superhamiltonian cycle from superhamiltonian cycles for the pieces. The key is the following definition. Let G be an outerplanar graph, and let (x, y) be a separating edge for G. A set of edges $P \subset E$ is *parallel* to (x, y) if

(1) $(x, y) \in P$;

(2) if $(u, v), (w, z) \in P$, then $\{u, v\} \cap \{w, z\} = \emptyset$ (there are no shared endpoints);

(3) P can be ordered as $\{(u_1, v_1), (u_2, v_2), \cdots, (u_k, v_k)\}$ in such a way that

$$u_1 < u_2 < \cdots < u_k < v_k < \cdots < v_2 < v_1$$

(the edges of P nest).

A sample set of parallel edges for a graph G is shown in Fig. 15 by dashed lines. A set P of parallel edges is *maximal* if no edge of G can be added to P to obtain a larger set of parallel edges.

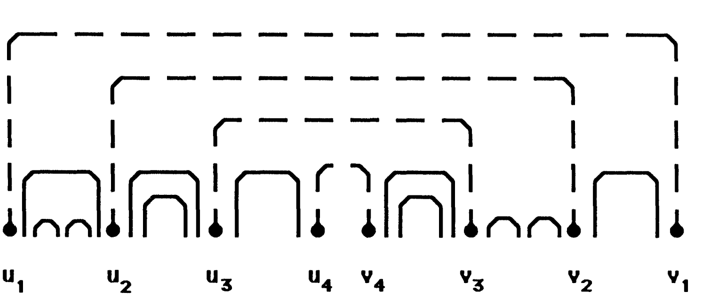


FIG. 15. *Parallel edges in G.*

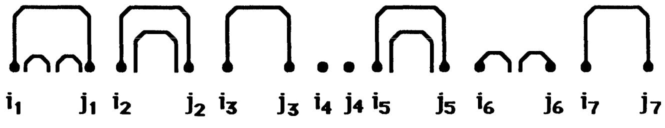


FIG. 16. *Removal of V_P.*

Suppose P is a maximal set of parallel edges for G containing the separating edge (x, y). Let V_P be the set of endpoints of edges in P. The removal of the vertices V_P from G separates the interval $[1, n]$ into some number of subintervals. Let G_P be the subgraph of G resulting from the removal of V_P and all incident edges. Let $[i_1, j_1], \cdots, [i_s, j_s]$ be these subintervals in left-to-right order. The planarity of G and the maximality of P guarantees that there is no edge of G between two vertices in different subintervals. This in turn guarantees that G_P can be obtained an alternate way: G_P is the (disjoint) union of the induced subgraphs $G[i_k, j_k]$, $1 \leqq k \leqq s$. By Lemma 3, the presence of a separating edge in P guarantees that

$$\text{size } \{[i_k, j_k]\} < \tfrac{2}{3}n, \qquad 1 \leqq k \leqq s.$$

Figure 16 shows the graph of Fig. 15 after the removal of V_P.

The algorithm is applied recursively to each $G[i_k, j_k]$ to obtain a superhamiltonian cycle for each. To obtain a superhamiltonian cycle for G, one need only reintroduce the endpoints of the parallel edges V_P. A second look at Fig. 15 provides inspiration. If each subinterval $[i_s, j_s]$ were replaced by an edge $(i_s - 1, j_s + 1)$ between two vertices in V_P, the result is the one-page embedding of a ladder where *all* the rungs nest. The construction of a superhamiltonian cycle H for G is patterned after the superhamiltonian cycle for a ladder, as illustrated in Fig. 6. Appropriately, we name the construction of H the *ladder construction.*

There are two cases to consider, depending on whether or not the edge $(1, n)$ is in P. The case $(1, n) \in P$ illustrates all the ideas and is simply modified to cover the case $(1, n) \not\in P$.

Start with the picture of the parallel edges alone in Fig. 17. Some lower edges are added to obtain a supercycle containing exactly the vertices in V_P. This supercycle is indicated in Fig. 18 by arrows. It remains to place all the subintervals within this supercycle. To accomplish this, each lower edge is replaced by new lower edges that connect two subintervals into its place in the supercycle. For a right arrow (u_k, u_{k+1}), the result is as in Fig. 19. For a left arrow (v_{k-1}, v_k), the result is as in Fig. 20; t is chosen so that $[i_t, j_t]$ is the subinterval between v_{k+1} and v_k.

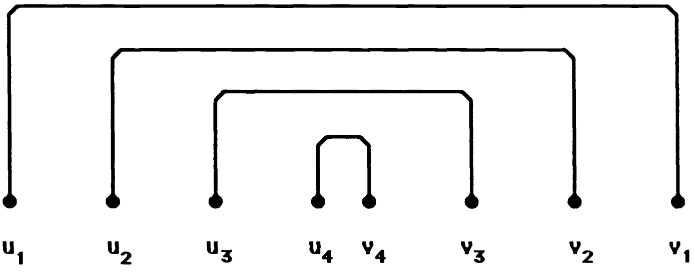


FIG. 17. *Parallel edges.*

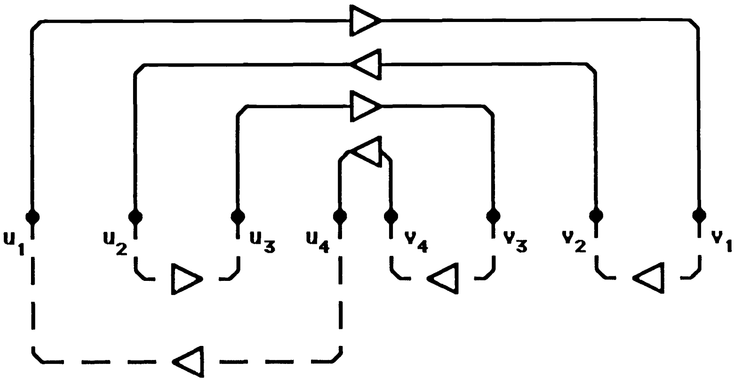


FIG. 18. *Supercycle for parallel edges.*

For the case $(1, n) \not\in P$, $[i_1, j_1]$ is to the left of the ladder and $[i_s, j_s]$ is to the right of the ladder. The connection of $[i_1, j_1]$ into H is shown in Fig. 21. The connection of $[i_s, j_s]$ into H is shown in Fig. 22.

5. The algorithm. This section describes our algorithm for embedding a d-valent outerplanar graph in a two-page book with logarithmic pagewidth. The correctness of the algorithm, embodied in Theorem 4, is given in the next section. Section 8 analyzes the performance of the algorithm.

For the statement of our algorithm, see Algorithm 1. As the algorithm is recursive, it is useful to give it a name. The name is TRADEOFF. TRADEOFF is a recursive function which has as input the d-valent outerplanar graph G and as output the planar supergraph G' having hamiltonian cycle H and vertices of attachment x and y.

It is to be noted that, for simplicity, certain trivial cases are not included in the statement of TRADEOFF. These cases occur when a recursive invocation of TRADEOFF returns an empty G'. This cannot occur in step 5, as each subinterval contains at least one vertex. However, it can occur in step 9 when some G_k is empty. In that case, the ladder construction merely skips the empty interval $[i_k, j_k]$ (which is caused by two adjacent elements of V_P).

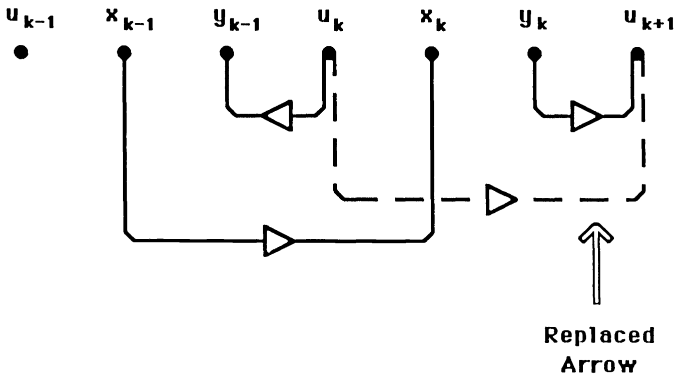


FIG. 19. *Replacing a right lower edge.*

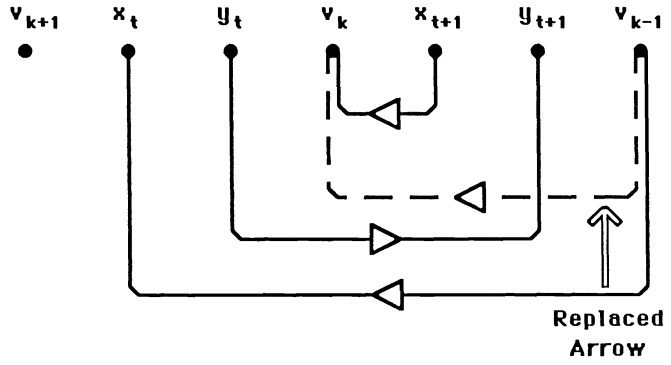


FIG. 20. *Replacing a left lower edge.*

ALGORITHM 1. *The Tradeoff Algorithm.*

Function TRADEOFF (G), **returns** (G', H, (x, y)).

```
(1) (Trivial cases)
    If $G = \emptyset$, then assign $G' = \emptyset$, $H = \emptyset$ and $(x, y) =$ *undefined.*
    If $V = \{1\}$, then assign $G' = (\{1\}, \{(1, 1)\})$, $H = \{(1, 1)\}$ and $(x, y) = (1, 1)$.
    If $V = \{1, 2\}$, then assign $G' = (\{1, 2\}, \{(1, 2), (1, 2)\})$, $H = \{(1, 2), (1, 2)\}$ and $(x, y) = (1, 2)$.
    Return $(G', H, (x, y))$.
(2) Let $S = \{m_j | 1 \leqq j \leqq q\}$ be the set of exposed vertices of $G$ in increasing order.
(3) Choose $k$, $1 \leqq k \leqq q - 1$ such that $B[m_k, m_{k+1}]$ is the key block of $G$.
(4) If $B[m_k, m_{k+1}] = G$, then go to step 7.

                    String Construction

(5) ($G$ has more than one block.) For $1 \leqq j < k$, assign
        $(G'_j, H_j, (x_j, y_j)) = \text{TRADEOFF } (G[m_j, m_{j+1}))$.
    For $j = k$, assign
        $(G'_j, H_j, (x_j, y_j)) = \text{TRADEOFF } (G[m_k, m_{k+1}])$.
    For $k < j < q$, assign
        $(G'_j, H_j, (x_j, y_j)) = \text{TRADEOFF } (G(m_j, m_{j+1}])$.
```

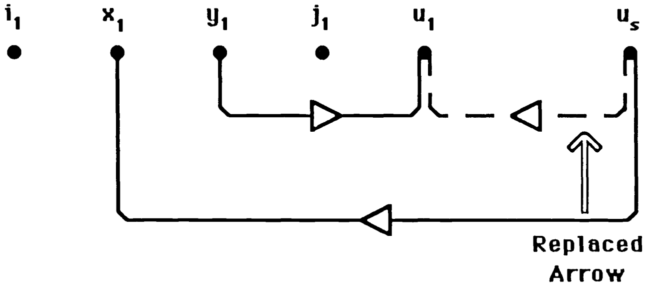


FIG. 21. *Adding a subinterval on the left.*

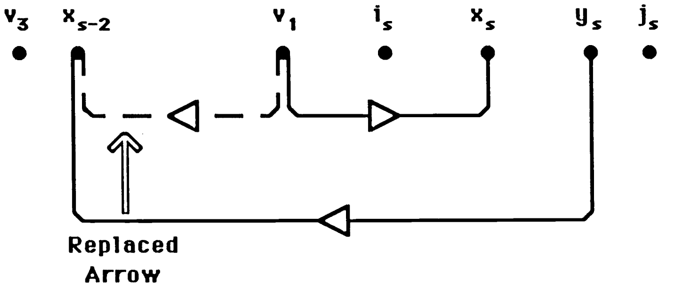


FIG. 22. *Adding a subinterval on the right.*

(6) Use the string construction to obtain G', H and (x, y). Return $(G', H, (x_1, y_q))$.

Ladder Construction

(7) ($B[m_k, m_{k+1}] = G$.) Choose a separating edge (u, v) for G. If (u, v) is not already an edge of G, then add it as an upper edge.

(8) Choose P a maximal set of edges parallel to (u, v). Let V_P be the set of endpoints of edges in P.

(9) $V - V_P$ determines a sequence of disjoint subintervals $[i_1, j_1]$, $[i_2, j_2]$, $\cdots$, $[i_s, j_s]$. For $1 \leqq k \leqq s$, make the assignment:

$$(G'_k, H_k, (x_k, y_k)) = \text{TRADEOFF}\ (G[i_k, j_k]).$$

Construct G' from G and $\{G'_k\}$ using the ladder construction. Return $(G', H, (x, y))$.

We now describe TRADEOFF step by step.

(1) These are the trivial cases when $n \leqq 2$. If G is empty, return $G' = \emptyset$. If G is a single vertex, return G' having a single loop. If G has two vertices, then it has one edge (1, 2). Return G' having the added lower edge (1, 2) which is distinct from the upper edge (1, 2).

(2) From the one-page embedding of G, determine the exposed vertices of G. It is straightforward to accomplish this step in linear time; see Algorithm 2. Algorithm 2 requires time $O(dn)$ and generates the elements of S in increasing order.

(3) Choose the key block of G, $B[m_k, m_{k+1}]$. Clearly, this can be accomplished in time linear in $|S|$.

(4) This step determines which of two cases is current. If G is a single block, then the ladder construction is applied (steps 7 through 9). If G has more than one block, then the string construction is applied (steps 5 and 6).

(5) Decompose the interval $[1, n]$ into subintervals so that the key block $B[m_k, m_{k+1}]$ is one of the subintervals. Note that each side subinterval contains fewer than $\frac{1}{2}n$ vertices. Apply TRADEOFF to the graphs induced by each subinterval to obtain supergraphs G'_j, $1 \leqq j \leqq q - 1$.

ALGORITHM 2. *Determining exposed vertices in linear time.*

```
(1)   Assign $S = \{1\}$ and $i = 1$.
(2)   If $i > n$, then halt.
(3)   Assign $i = \max\ \{i + 1, \max_{(i,k) \in E} k\}$.
(4)   Assign $S = S \cup \{i\}$. Go to step 2.
```

(6) Apply the string construction to obtain the planar hamiltonian supergraph G' and the hamiltonian cycle H for G'. Assign $(x, y) = (x_1, y_q)$. Return $(G', H, (x, y))$.

(7) We know that G is entirely covered by the edge $(1, n)$. We show that it is then safe to add a separating edge to G. If this is the initial call to TRADEOFF, we can always add a separating edge. If this is a deeper recursive call to TRADEOFF, we imagine that there are intervals to the left and right of $[1, n]$ with dangling edges incident to vertices in $[1, n]$. Since these dangling edges can only be incident to exposed vertices (in this case, 1 and n), any upper edge added to G at this recursive level cannot cross an edge at a higher recursive level. The determination of a suitable separating edge is accomplished in linear time by Algorithm 3. (Note that the triangulated G has a linear number of edges.)

(8) Select a maximal set of parallel edges. The construction of P is accomplished in linear time by Algorithm 4.

(9) This step completes the ladder construction. TRADEOFF is invoked recursively for each subinterval disjoint from V_P. G' and H are obtained by the ladder construction described in the previous section.

ALGORITHM 3. *Finding a separating edge.*

```
(1)   Triangulate the interior faces of $G$.
(2)   Examine each edge $(u, v)$ of the triangulated $G$ to find one such that
      $\frac{1}{3}n \leqq (v - u) \leqq \frac{2}{3}n$.
```

ALGORITHM 4. *Generating a maximal set of parallel edges.*

```
(1)   Assign $P = \{(u, v)\}$, $s = u - 1$ and $t = v$.
(2)   If $s < 1$, then go to step 4.
(3)   Assign $r = \max\{1, \max_{(s,k)\in E} k\}$. If $r \leqq t$, then assign $s = s - 1$ and go to
      step 2. Else assign $P = P \cup \{(s, r)\}$, $s = s - 1$ and $t = r$ and go to step 2.
(4)   Assign $s = u + 1$, and $t = v$.
(5)   If $s \geqq t$, then halt.
(6)   Assign $r = \min\{n, \min_{(s,k)\in E} k\}$. If $r < t$, then assign $P = P \cup \{(s, r)\}$,
      $s = s + 1$ and $t = r$ and go to step 5. Else, assign $s = s + 1$ and go to
      step 5.
```

6. Correctness. In this section, we demonstrate the correctness of algorithm TRADEOFF via the following theorem.

THEOREM 4. *Let G be a d-valent outerplanar graph. Let $(G', H, (x, y))$ result from applying TRADEOFF to G. Then H is a superhamiltonian cycle for G with the following property: following H from x to y yields a two-page embedding of G with pagewidth* $<Cd \log n$, *where C is a constant that can be chosen to have any value* $\geqq 8/(\log 3/2)$.

Proof. The proof decomposes naturally into the proof of pagenumber (Lemma 5) and the proof of pagewidth (Lemma 6). □

LEMMA 5. *Given the assumptions of Theorem* 4, *following H from x to y yields a two-page embedding of G.*

Proof. The proof is by induction on n. The inductive hypothesis is

(H.1) $G \subset G'$;

(H.2) G' is planar;

(H.3) H is a hamiltonian cycle of G';

(H.4) $(x, y) \in H$ is a lower edge of G' such that there is no lower edge (u, v) of

G' with $u < x \leqq y < v$ (i.e., x and y are on the unbounded region of the lower half-plane).

Step 1 of Algorithm 1 guarantees that the inductive hypothesis is satisfied when $n \leqq 2$.

For purposes of induction, assume that the inductive hypothesis is true for graphs of size less than n and that $n > 2$. There are two cases determined by the cardinality of the set S of exposed vertices of G: (1) $|S| > 2$ and (2) $|S| = 2$.

(1) $|S| > 2$. TRADEOFF applies the string construction in steps 5 and 6. The inductive hypothesis guarantees that after the applications of TRADEOFF to all the subintervals, each x_j and each y_j is on the unbounded region of the lower half-plane. Therefore, the lower edges (y_j, x_{j+1}), $1 \leqq j \leqq q - 1$ and (x_1, y_q) can be added while maintaining planarity (H.2). Clearly, $G \subset G'$ (H.1), and H is a hamiltonian cycle of G' (H.3). Finally $x = x_1$ and $y = y_q$ satisfy (H.4).

(2) $|S| = 2$. The edge $(1, n)$ is in G and covers all other upper edges. TRADEOFF applies the ladder construction to G in steps 7 through 9. In § 2.2, the addition of the separating upper edge (u, v) was shown to maintain planarity. In step 9, the application of TRADEOFF to each $G[i_k, j_k]$ yields $(G'_k, H, (x_k, y_k))$ that satisfies the inductive hypothesis. In particular, (H.4) applies to each (x_k, y_k). Since each x_k and y_k is on the unbounded region of the lower half-plane, the ladder construction yields a planar result (H.2). The ladder construction also makes $G \subset G'$ (H.1) and H a hamiltonian cycle of G' (H.3). Finally, (x, y) is explicitly chosen to satisfy (H.4).

This extends the induction for arbitrary G. Since H is a hamiltonian cycle of a planar supergraph of G, it yields a two-page embedding of G [1]. □

To complete the proof of Theorem 4, we must bound the pagewidth of the two-page embedding. It is sufficient to bound the cutwidth of the underlying linear embedding. We use the notation cw (H) to mean the cutwidth of the linear embedding obtained by following H from x through y. (G, x and y will be clear from context.) If i and j are vertices in H such that i comes before j in the linear embedding, define cw $([i, j])$ to be the cutwidth of the linear subembedding from i to j.

LEMMA 6. *Given the assumptions of Theorem* 4, $\mathrm{cw}(H) < Cd \log n$, *where* $C = 8/(\log 3/2)$.

Proof. The proof is by induction on n. The statement of the inductive hypothesis mirrors the two cases of the algorithm. The inductive hypothesis is

(I.1) If G has more than one block, then

$$\mathrm{cw}\,(H) < Cd \log n;$$

(I.2) If G is a single block, then

$$\mathrm{cw}\,(H) \leqq \max\,(1, Cd \log n) - \mathrm{def}\,\{[1, n]\}.$$

Some explanation of the presence of the edge deficit in (I.2) is in order. In the string construction, a large key block $[m_k, m_{k+1}]$ must be able to absorb def $\{[m_k, m_{k+1}]\}$ additional cutwidth, as its cutwidth will dominate the cutwidth of the entire string construction. The precise meaning of this statement will be clear from the proof. The max $(1, Cd \log n)$ takes care of the case $n = 1$. Note that a G with a single vertex can never be the key block in a string construction.

For the basis of the induction, it is easy to check the inductive hypothesis for $n = 1$ and $n = 2$.

For purposes of induction, assume that the inductive hypothesis is true for graphs of size less than n and that $n > 2$. There are two cases: (1) G has more than one block and (2) G is a single block.

(1) *G has more than one block.* In this case, the string construction is applied (steps 5 and 6). Let us examine the linear order induced by H and (x, y) on V. H_1 is a superhamiltonian cycle for $G([m_1, m_2))$ that begins at $x = x_1$ and ends at y_1. As such, H_1 can be viewed as a permutation on $[m_1, m_2)$. The string construction places the vertices of $[m_1, m_2)$ first in H, in this permuted order. Similarly, the vertices of $[m_2, m_3)$ come next in H, in the permuted order given by H_2. In general, the $q - 1$ subintervals appear in the same order in H as they do in the partition, though H permutes the vertices within each subinterval. The permutation of the jth subinterval is always that of H_j.

It is now possible to bound cw (H) based on $\{\text{cw}\,(H_j)\}$. First, consider the cutwidth of H between two subintervals, that is, cw $([y_j, x_{j+1}])$, $1 \leqq j \leqq q - 2$. Suppose $j < k$. Then the only edges that pass over the interval $[y_j, x_{j+1}]$ are dangling edges from m_{j+1} to $[m_j, m_{j+1})$. Hence,

$$\text{cw}\,([y_j, x_{j+1}]) \leqq d - 1 < Cd \log n.$$

If $j \geqq k$, by a similar argument, we have

$$\text{cw}\,([y_j, x_{j+1}]) \leqq d - 1 < Cd \log n.$$

Second, consider the cutwidth over a side subinterval. Consider the jth subinterval in H, $[x_j, y_j]$. If $j < k$, then there are at most $(d - 1)$ dangling edges from m_{j+1} to $[m_j, m_{j+1})$ that can contribute to cw $([x_j, y_j])$ and at most d dangling edges from m_j to $[m_{j-1}, m_j)$ that can contribute to cw $([x_j, y_j])$. Hence by (I.1)

$$\begin{aligned}\text{cw}\,([x_j, y_j]) &< 2d + Cd \log\,(\text{size}\,\{[m_j, m_{j+1})\})\\ &< Cd \log n\end{aligned}$$

since size $\{[m_j, m_{j+1})\} < \frac{1}{2}n$. If $j > k$, we have similarly

$$\begin{aligned}\text{cw}\,([x_j, y_j]) &< 2d + Cd \log\,(\text{size}\,\{(m_j, m_{j+1}]\})\\ &< Cd \log n.\end{aligned}$$

Third and finally, consider the cutwidth over the key block, $B[m_k, m_{k+1}]$. By (I.2),

$$\text{cw}\,([x_k, y_k]) \leqq (Cd \log\,(\text{size}\,\{[m_k, m_{k+1}]\})) - \text{def}\,\{[m_k, m_{k+1}]\}.$$

The only dangling edges that can contribute to the cutwidth over $[x_k, y_k]$ are those incident to m_k and m_{k+1}. There are def $\{[m_k, m_{k+1}]\}$ of these. Hence

$$\begin{aligned}\text{cw}\,([x_k, y_k]) &\leqq Cd \log\,(\text{size}\,\{[m_k, m_{k+1}]\})\\ &< Cd \log n.\end{aligned}$$

Putting these three results together yields cw $(H) < Cd \log n$. Thus G satisfies (I.1).

(2) *G is a single block.* In this case, the ladder construction is applied (steps 7 through 9). The subintervals are $[i_1, j_1], \cdots, [i_s, j_s]$. Let

$$P = \{(u_1, v_1), (u_2, v_2), \cdots, (u_t, v_t)\}$$

where

$$u_1 < u_2 < \cdots < u_t < v_t < \cdots < v_2 < v_1 \quad \text{and} \quad t = \lfloor (s+1)/2 \rfloor.$$

We first consider the case $(1, n) \in P$. We can represent the order in H of the vertices of V_P and of the subintervals by the following string:

$$u_1 v_1 [i_{s-1}, j_{s-1}][i_s, j_s] v_2 u_2 [i_1, j_1][i_2, j_2] u_3 v_3 [i_{s-3}, j_{s-3}][i_{s-2}, j_{s-2}] v_4 u_4 [i_3, j_3][i_4, j_4] u_5 v_5 \cdots.$$

Of course, the vertices of the subintervals are permuted with H as they were in case (1).

From the ladder construction, there are four recognizable *types* of subintervals, two types on the left and two types on the right. While we could write down subscript formulas for each of the four types, for the cutwidth argument it is sufficient to consider the following four representatives of the four types: $[i_3, j_3]$, $[i_4, j_4]$, $[i_{s-3}, j_{s-3}]$ and $[i_{s-2}, j_{s-2}]$. The only edges that add to cw (H_k) are edges incident to vertices in V_P that pass over the kth subinterval in H. The diagram in Fig. 23 illustrates the *potential* for a vertex in V_P to have edges incident to some subinterval. For example, u_2 or v_2 might have one or more edges to subintervals $[i_1, j_1]$, $[i_s, j_s]$, $[i_2, j_2]$, and $[i_{s-1}, j_{s-1}]$. Since we are interested only in an upper bound on cutwidth, we ignore the possibility that the existence of some edge may preclude the existence of other edges.

We start with the type represented by subinterval $[i_3, j_3]$. An examination of the string for H together with Fig. 23 reveals the potential for edges passing over $[x_3, y_3]$ from u_3, v_3, u_4, v_4, u_5 and v_5 only. Hence, by inductive hypothesis,

$$\begin{aligned}\text{cw}\,([x_3, y_3]) &\leqq 6d + \text{cw}\,(H_3)\\ &\leqq 6d + Cd \log\,(\text{size}\,\{[i_3, j_3]\})\\ &\leqq 6d + Cd \log \tfrac{2}{3}n\\ &\leqq (Cd \log n) - 2d\\ &< (Cd \log n) - \text{def}\,\{[1, n]\}\end{aligned}$$

since each subinterval contains at most $\frac{2}{3}n$ vertices.

Similarly, consideration of the three types represented by $[i_4, j_4]$, $[i_{s-3}, j_{s-3}]$ and $[i_{s-2}, j_{s-2}]$ reveals that at most 6 vertices in V_P can have incident edges adding to the cutwidth of a subinterval. Hence, for all subintervals $[x_k, y_k]$ in H,

$$\text{cw}\,([x_k, y_k]) \leqq (Cd \log n) - \text{def}\,\{[1, n]\}.$$

Consideration of intervals in H between the subintervals (e.g., $[u_5, v_5]$) yields no worse an upper bound. Hence we conclude that cw $(H) \leqq (Cd \log n) - \text{def}\,\{[1, n]\}$.

The case in which $(1, n) \notin P$ is similar to the preceding case. The additional left or right subinterval cannot boost the cutwidth above $(Cd \log n) - \text{def}\,\{[1, n]\}$. Hence in all cases, (I.2) holds.

This completes the induction and the proof of the lemma. □

7. Performance. In this section, we analyze the time and space complexity of TRADEOFF. Of course, the complexity depends on the representation of data. While

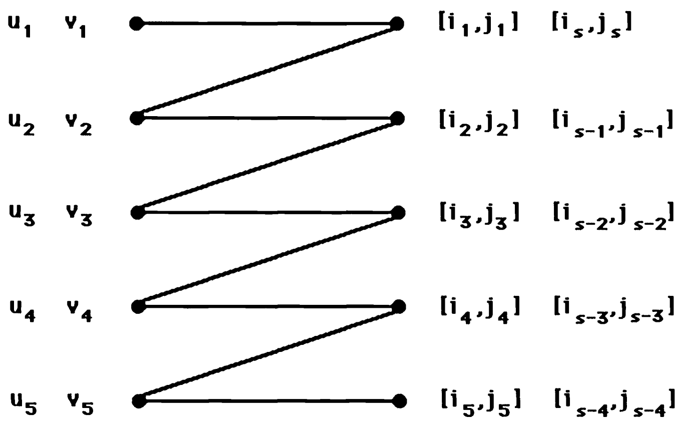

u_1 v_1
u_2 v_2
u_3 v_3
u_4 v_4
u_5 v_5
$[i_1, j_1]$ $[i_s, j_s]$
$[i_2, j_2]$ $[i_{s-1}, j_{s-1}]$
$[i_3, j_3]$ $[i_{s-2}, j_{s-2}]$
$[i_4, j_4]$ $[i_{s-3}, j_{s-3}]$
$[i_5, j_5]$ $[i_{s-4}, j_{s-4}]$

FIG. 23. *Proof of Lemma* 9.

we do not prescribe the details of the data representation, we do require that the representation make elementary operations efficient (i.e., constant time per edge or vertex). A place where this requirement is crucial is Algorithm 3 for finding a separating edge. To accomplish step 1 in linear time, it is necessary to be able to recognize the next (counterclockwise) edge of an interior face in constant time. It is easy to represent G so that this is possible. A reasonable representation puts all the edges adjacent to a vertex of G in a circular list in counterclockwise order; for each edge (u, v), there is a link from its position in the u-list to its position in the v-list. In this representation, the edges of an interior face can be traversed in constant time per edge.

First, we note that all operations of TRADEOFF performed on G except the recursive calls require linear time. From the description of the steps in § 3, all steps are clearly linear time except steps 6 and 9. From the description of the string construction, G' can be constructed in linear time from the $\{G_j\}$ (step 6). Similarly, the ladder construction can be accomplished in linear time (step 9). Hence, the entire algorithm excluding recursive calls can be implemented in linear time.

Let $T(n)$ be the time complexity of TRADEOFF. Let $n_1, n_2, \cdots, n_p$ be the sizes of the subintervals either in step 5 or in step 9, depending on which case holds. Then, $\sum_{k=1}^{p} n_k \leqq n$ and $n_k \leqq \frac{2}{3}n$, $1 \leqq k \leqq p$. By the result of the previous paragraph, there exists a constant c such that

$$T(n) \leqq cn + \sum_{k=1}^{p} T(n_k).$$

LEMMA 7. *If $T(1)$ is one unit of time, then for all $n > 1$,*

$$T(n) \leqq (c/\log \tfrac{2}{3})n \log n.$$

Proof. By induction on n. The lemma is certainly true for $n = 2$. Assume $n > 2$ and assume the truth of the lemma for values smaller than n. Then,

$$\begin{aligned} T(n) &= cn + \sum_{k=1}^{p} T(n_k) \\ &\leqq cn + \sum_{k=1}^{p} \left(c \Big/ \log \frac{3}{2}\right) n_k \log n_k \\ &\leqq cn + \left(c \Big/ \log \frac{3}{2}\right) \sum_{k=1}^{p} n_k \log \frac{2}{3} n \\ &= cn + \left(c \Big/ \log \frac{3}{2}\right) n \log \frac{2}{3} n \\ &= cn + \left(c \Big/ \log \frac{3}{2}\right) n \log n - \left(c \Big/ \log \frac{3}{2}\right) n \log \frac{3}{2} \\ &= \left(c \Big/ \log \frac{3}{2}\right) n \log n. \end{aligned}$$

The lemma follows by induction. □

The space requirements of TRADEOFF are clearly n times some small constant. We thus have the following.

THEOREM 8. TRADEOFF *has time complexity at most $C_1 n \log n$ and space complexity at most $C_2 n$, for small constants C_1, C_2.*

8. Conclusion. We have investigated tradeoffs between pagenumber and pagewidth that are significant in a VLSI context. Our main result is an algorithm for obtaining a book embedding for outerplanar graphs that is within a constant factor of optimal in VLSI area for the class of outerplanar graphs. While this near-optimality is not guaranteed for *individual* outerplanar graphs, we know of no example of an outerplanar graph for which our algorithm fails to obtain near-optimal area. Our algorithm embeds any d-valent n-vertex outerplanar graph in a two-page book with at most $Cd \log n$ pagewidth, $C = 8/(\log 3/2)$; the algorithm executes in time $O(n \log n)$. We show that at the cost of one additional page above optimal pagenumber, layouts of near-optimal cutwidth for outerplanar graphs can be obtained constructively.

Our result is applicable to the motivating DIOGENES design problem. The result bounds the area of a two-stack DIOGENES layout for a circuit represented by an outerplanar graph.

A fruitful area for further research is tradeoffs between pagenumber and pagewidth. It is not known how prevalent such tradeoffs are or whether dramatic tradeoffs exist for any pagenumber. In the context of VLSI problems, algorithms for embedding a graph in a bounded number of pages with pagewidth close to the cutwidth of the graph could be most practical. We do not know of any example where adding one or two pages above the pagenumber of G does not give us an embedding whose pagewidth is within a small constant factor of the pagewidth of G; there is hope that such algorithms exist for important classes of graphs. In particular, we believe that such an algorithm is possible for planar graphs. The algorithm would embed any d-valent planar graph in a B-page book with $Cd\sqrt{n}$ pagewidth where B and C are small constants. Our planar graph algorithm [4] or the similar algorithm of Yannakakis [11] might serve as the starting point for obtaining bounded pagenumber. The small pagewidth would depend on a version of Lipton and Tarjan's [7] planar separator theorem tailored to this problem. The result of Miller [8] on separating *cycles* in planar graphs is relevant here, though it is not sufficient by itself.

Acknowledgments. The author wishes to thank Arnold Rosenberg for suggesting the problem and for helpful discussions. He also wishes to thank the referee whose detailed report improved the presentation of the paper greatly.

REFERENCES

[1] F. Bernhart and P. C. Kainen, *The book thickness of a graph*, J. Combin. Theory Ser. B, 27 (1979), pp. 320–331.

[2] F. R. K. Chung, F. T. Leighton and A. L. Rosenberg, *Embedding graphs in books: a layout problem with applications to VLSI design*, this Journal, 8 (1987), pp. 33–58.

[3] M. R. Garey, D. S. Johnson, G. L. Miller and C. H. Papadimitriou, *The complexity of coloring circular arcs and chords*, this Journal, 1 (1980), pp. 216–227.

[4] L. S. Heath, *Embedding planar graphs in seven pages*, 25th IEEE Symposium on Foundations of Computer Science, 1984, pp. 74–83.

[5] ———, *Algorithms for embedding graphs in books*, Ph.D. dissertation. Univ. of North Carolina at Chapel Hill, Dept. of Computer Science Technical Report TR 85-028, 1985.

[6] T. Lengauer, *Upper and lower bounds on the complexity of the min-cut linear arrangement problem on trees*, this Journal, 3 (1982), pp. 99–113.

[7] R. J. Lipton and R. E. Tarjan, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–189.

[8] G. L. Miller, *Finding small simple cycle separators for 2-connected planar graphs*, 16th ACM Symposium on Theory of Computing, 1984, pp. 376–382.

[9] A. L. Rosenberg, *The DIOGENES approach to testable fault-tolerant arrays of processors*, IEEE Trans. Comput., C-32 (1983), pp. 902–910.

[10] M. M. Syslo, *Characterizations of outerplanar graphs*, Discrete Math., 26 (1979), pp. 47–53.

[11] M. Yannakakis, *Four pages are necessary and sufficient for planar graphs*, 18th ACM Symposium on Theory of Computing, 1986, pp. 104–108.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

A COMPLEX ORTHOGONAL-SYMMETRIC ANALOG OF THE POLAR DECOMPOSITION*

DIPA CHOUDHURY† AND ROGER A. HORN‡

Abstract. Every square complex matrix A can be factorized as $A = UH$, where U is unitary and H is positive semi-definite Hermitian. If A is nonsingular, it is known that one may write $A = QS$, where Q is complex orthogonal and S is complex symmetric. We develop necessary and sufficient conditions for there to be a factorization of this type when A is singular, and we give sufficient conditions for there to be at least one such factorization in which the factors commute.

Key words. complex orthogonal matrix, complex symmetric matrix, polar decomposition, matrix square root

AMS(MOS) subject classifications. 15A23, 15A57

Let $M_{m,n}$ denote the set of complex m-by-n matrices and write $M_n \equiv M_{n,n}$. We shall use Q to denote a complex orthogonal matrix ($Q \in M_n$, $QQ^T = I$) and S to denote a complex symmetric matrix ($S \in M_n$, $S = S^T$). A set of vectors $\{x_1, x_2, \cdots, x_n\} \subset \mathbb{C}^n$ is said to be *rectangular* if $x_i^T x_j = 0$ whenever $i \neq j$; it is *rectanormal* if it is rectangular and $x_i^T x_i = 1$ for all $i = 1, 2, \cdots, k$. A vector $x \in \mathbb{C}^n$ such that $x^T x = 0$ is said to be *isotropic*; the vector $x = [1\ i]^T \in \mathbb{C}^2$ is an example of a nonzero isotropic vector. Every subspace of $\mathbb{C}^n$ has a rectangular basis, and every rectangular basis of a given subspace contains the same number of isotropic vectors. A subspace is said to be *nonsingular* if it has a rectanormal basis. For basic facts about rectangular sets and the geometry associated with the bilinear form $b(x, y) = y^T x$, see [1], [7], [8].

It is known that every nonsingular complex matrix $A \in M_n$ can be written in the form $A = QS$ where S is symmetric and Q is complex orthogonal [4, Vol. II, Thm. 3 of Chap. XI], but this factorization may not be possible if A is singular. For example, if

$$A = \begin{bmatrix} 1 & i \\ 0 & 0 \end{bmatrix}$$

could be factored as $A = QS$, then $A^T A = S^T Q^T Q S = S^2$ has a (symmetric) square root, but the Jordan canonical form of

$$A^T A = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix} \quad \text{is} \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

and hence $A^T A$ does not have a square root. We develop necessary and sufficient conditions for a possibly singular matrix to be factored as $A = QS$.

If $A \in M_n$ can be expressed as $A = QS$, then $A = QS = QSQ^T Q = \hat{S}Q$, where $\hat{S} \equiv QSQ^T$ is a complex symmetric matrix. Similarly, if $A \in M_n$ can be factored as $A = SQ$, then $A = SQ = QQ^T SQ = Q\tilde{S}$, where $\tilde{S} \equiv Q^T SQ$ is a complex symmetric matrix. Thus, the existence of either factorization $A = QS$ or $A = SQ$ implies the existence of the other. Here we consider the factorization $A = QS$.

Our discussion uses a special case of a general criterion that tells when two given matrices $X_1, X_2 \in M_{n,k}$, $k \leqq n$, are orthogonal transforms of each other, i.e., when there is an orthogonal $Q \in M_n$ such that $X_1 = QX_2$.

* Received by the editors January 29, 1986; accepted for publication (in revised form) August 21, 1986.

† Department of Mathematical Sciences, Loyola College, Baltimore, Maryland 21218.

‡ Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218.

LEMMA 1. *Let $X_1, X_2 \in M_{n,k}$ with $1 \leqq k \leqq n$ and let rank $X_1 = q$. There is a complex orthogonal $Q \in M_n$ such that $X_1 = QX_2$ if and only if the following two conditions are satisfied*:

(a) $X_1^T X_1 = X_2^T X_2$;

(b) rank X_1 = rank X_2; *in addition, if* $1 \leqq q < k$, *let* $P \in M_k$ *be a permutation matrix such that* $X_1 P = [\hat{X}_1 \,\vert\, \tilde{X}_1]$, $\hat{X}_1 \in M_{n,q}$ *has full rank* q, *and* $\tilde{X}_1 = \hat{X}_1 C$ *for some* $C \in M_{q,k-q}$. *Then* $X_2 P = [\hat{X}_2 \,\vert\, \hat{X}_2 C]$, *where* $\hat{X}_2 \in M_{n,q}$.

Condition (b) *is equivalent to*

(b′) *There is a nonsingular* $B \in M_n$ *such that* $X_1 = BX_2$.

Proof. Let $X_1, X_2 \in M_{n,k}$ with $1 \leqq k \leqq n$ and let rank $X_1 = q$. We first show that (b) and (b′) are equivalent.

If $X_1 = BX_2$ for some nonsingular $B \in M_n$, then X_1 and X_2 must have the same rank. If $P \in M_k$ is a permutation matrix such that $X_1 P = [\hat{X}_1 \,\vert\, \hat{X}_1 C]$ for some $\hat{X}_1 \in M_{n,q}$ with full rank q and $C \in M_{q,k-q}$, partition $X_2 P = [\hat{X}_2 \,\vert\, \tilde{X}_2]$ with $\hat{X}_2 \in M_{n,q}$. Then $[\hat{X}_1 \,\vert\, \hat{\mathrm{X}}_1 C] = X_1 P = BX_2 P = [B\hat{X}_2 \,\vert\, B\tilde{X}_2]$, so $\hat{X}_1 = B\hat{X}_2$ and $B\tilde{X}_2 = \hat{X}_1 C = B\hat{X}_2 C$, which implies that $\tilde{X}_2 = \hat{X}_2 C$ since B is nonsingular. This shows that (b′) implies (b).

Conversely, assume (b) and let $X_1 P = [\hat{X}_1 \,\vert\, \hat{X}_1 C]$, $X_2 P = [\hat{X}_2 \,\vert\, \hat{X}_2 C]$ with $\hat{X}_1, \hat{X}_2 \in M_{n,q}$ both of full rank q. Then there are full rank matrices Y_1, $Y_2 \in M_{n,n-q}$ such that $Z_1 \equiv [\hat{X}_1 \,\vert\, Y_1]$ and $Z_2 \equiv [\hat{X}_2 \,\vert\, Y_2]$ are both nonsingular (just extend the columns of $\hat{X}_1$ and $\hat{X}_2$ to bases of $\mathbb{C}^n$). If we write

$$R \equiv \begin{bmatrix} I & C \\ 0 & 0 \end{bmatrix} \in M_n,$$

then $X_1 P = Z_1 R$, $R = Z_1^{-1} X_1 P$, $X_2 P = Z_2 R = Z_2 Z_1^{-1} X_1 P$, *and* $X_1 = (Z_1 Z_2^{-1}) X_2$, which is condition (b) with $B \equiv Z_1 Z_2^{-1}$.

To prove the primary assertion of the theorem, notice that if $X_1 = QX_2$ for some complex orthogonal $Q \in M_n$, then (a) and (b′) follow immediately. Conversely, suppose (a) and (b′) are satisfied. Let V_i denote the span of the columns of X_i, $i = 1, 2$, let $X_i = [\xi_1^{(i)} \cdots \xi_k^{(i)}]$ be partitioned according to its columns for $i = 1, 2$, and consider the linear mapping T: $V_2 \to V_1$ given on the columns of X_2 by $T(\xi_j^{(2)}) \equiv \xi_j^{(1)}, j = 1, \cdots, k$. Then T is well defined, linear, one to one, and onto by (b′), and (a) guarantees that T is an isometry with respect to the bilinear form $b(x, y) \equiv y^T x$. Witt's Theorem [8, Thm. 202.1] ensures that T can be extended to an isometry of $\mathbb{C}^n$, whose representation in the standard orthonormal basis is a complex orthogonal matrix Q with $QX_2 = X_1$. □

If $A \in M_n$ can be expressed as $A = QS$, then for any permutation matrices P, $R \in M_n$, $PAR = PQSR = PQRR^T SR = \hat{Q}\hat{S}$, where $\hat{Q} \equiv PQR$ is a complex orthogonal matrix and $\hat{S} \equiv R^T SR$ is a complex symmetric matrix. Thus, in an effort to characterize those singular $A \in M_n$ that can be written as $A = QS$, there is no loss of generality if we assume that A is given in the partitioned form $A = [\hat{A} \,\vert\, \hat{A}C]$ where the columns of $\hat{A} \in M_{n,k}$ are linearly independent and $C \in M_{k,n-k}$. The only exceptional case is $A = 0$, which clearly has a factorization of the desired form.

THEOREM 2. *Let* $A \in M_n$ *be a given nonzero matrix with rank* $k \leqq n$, *and let* $P \in M_n$ *be a permutation matrix such that* $AP = [\hat{A} \,\vert\, \hat{A}C]$, *where* $\hat{A} \in M_{n,k}$ *has full rank and* $C \in M_{k,n-k}$. *There exists a complex orthogonal matrix* Q *and a complex symmetric matrix* S *such that* $A = QS$ *if and only if there exists a symmetric* $S \in M_n$ *such that*

(1) $S^2 = A^T A$, *and*

(2) $SP = [\hat{S} \,\vert\, \hat{S}C]$ *for some* $\hat{S} \in M_{n,k}$ *with full rank.*

Proof. Without loss of generality we assume that $A = [\hat{A} \,\vert\, \hat{A}C]$, where $\hat{A} \in M_{n,k}$ has full rank $k \geqq 1$. If $k = n$, we adopt the convention that the terms $\hat{A}C$ and $\hat{S}C$ are absent from the respective partitioned presentations of AP and SP, $P = I$, $A = \hat{A}$, and $S = \hat{S}$.

Suppose A can be written as $A = QS$. Then $A^TA = SQ^TQS = S^2$, and hence S is a symmetric square root of A^TA. Write $S = [\hat{S} \,|\, \tilde{S}]$, conformal with the indicated partition of A. Then $A = [\hat{A} \,|\, \hat{A}C] = QS = Q[\hat{S} \,|\, \tilde{S}] = [Q\hat{S} \,|\, Q\tilde{S}]$. Hence $\hat{A} = Q\hat{S}$ and $\hat{A}C = Q\tilde{S}$, i.e., $Q\hat{S}C = Q\tilde{S}$ and hence $\hat{S}C = \tilde{S}$ since Q is nonsingular. Thus, $S = [\hat{S} \,|\, \hat{S}C]$ and rank $\hat{S}$ = rank $Q\hat{S}$ = rank $\hat{A} = k$, so the stated conditions are necessary.

Conversely, suppose there is a symmetric $S \in M_n$ such that $S^2 = A^TA$ and $S = [\hat{S} \,|\, \hat{S}C]$ for some $\hat{S} \in M_n$ with full rank. Since $S^2 = S^TS$, conditions (a) and (b) in Lemma 1 are satisfied and there exists an orthogonal $Q \in M_n$ such that $A = QS$. □

Because of the equivalence of conditions (b) and (b′) in Lemma 1, condition (2) of the theorem is equivalent to

(2′) $A = BS$ for some nonsingular $B \in M_n$.

In the ordinary polar decomposition, we know that if $A = UH$ (where U is unitary and H is positive definite), then U commutes with H if and only if A is normal [5, Thm. 7.3.4]. Thus, the factors in any one polar decomposition of a given matrix $A \in M_n$ commute if and only if the factors in all polar decompositions of that matrix commute. What is the analogous situation for orthogonal-symmetric factorization?

If $A \in M_n$ and $A = QS$, where Q is complex orthogonal and S is symmetric, and if Q commutes with S then $AA^T = QSSQ^T = SQQ^TS = SQ^TQS = A^TA$. It is therefore necessary that $AA^T = A^TA$ if Q and S are to commute. If A is nonsingular, it is known [4, Vol. II, Thm. 3 of Chap. XI] that the condition $AA^T = A^TA$ is sufficient for there to be at least one factorization $A = QS$ in which Q commutes with S. If A is singular and $AA^T = A^TA$, however, it can happen that no matter how one forms $A = QS$, the factors will not commute.

Example. Let

$$A = \begin{bmatrix} 0 & i & -1 \\ -i & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Since A is skew-symmetric, $AA^T = A^TA$. This matrix can be factored as

$$A = \begin{bmatrix} 0 & i & -1 \\ -i & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -(1+i) & (1-i) \\ -(1+i) & (1-i) & 1 \\ (1-i) & 1 & (1+i) \end{bmatrix} \begin{bmatrix} 0 & -i & 1 \\ -i & (1-i) & (1+i) \\ 1 & (1+i) & (i-1) \end{bmatrix} \equiv QS.$$

One checks that $A^T = -A = SQ \neq QS = A$. This noncommutativity is no accident: there is no factorization of the form $A = QS$ in which the factors commute. Suppose $A = QS = SQ$, where

$$Q^T = \begin{bmatrix} a & b & c \\ e & f & g \\ h & k & m \end{bmatrix} \in M_3$$

is complex orthogonal and $S \in M_3$ is symmetric. Then $AQ^T = Q^TA = S$, which implies that

$$\begin{bmatrix} -bi+c & ai & -a \\ -fi+g & ei & -e \\ -ki+m & hi & -h \end{bmatrix} = Q^TA = S = AQ^T = \begin{bmatrix} ei-h & fi-k & gi-m \\ -ai & -bi & -ci \\ a & b & c \end{bmatrix}$$

and all of these matrices are symmetric. In particular, the (1, 2) entry of the first matrix equals the (2, 1) entry of the last matrix, i.e., $ai = -ai$ and hence $a = 0$. Also, the (2, 3) entry of the last matrix equals the (3, 2) entry of the same matrix, i.e., $b = -ci$. Therefore $a^2 + b^2 + c^2 = 0$. Thus, the first row of Q^T is isotropic, a contradiction of the assumption that Q is orthogonal. Observe that

$$A^TA = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & -i \\ 0 & -i & 1 \end{bmatrix},$$

so rank $A = 2 \neq 1 =$ rank A^TA. Notice that 0 is the only eigenvalue of A^TA, with geometric multiplicity 2, so A^TA is not diagonalizable.

For a given matrix $A \in M_n$, it is possible to have two different factorizations $A = QS$ and $A = \hat{Q}\hat{S}$, where $Q, \hat{Q} \in M_n$ are complex orthogonal and $S, \hat{S} \in M_n$ are symmetric, in which one pair of factors commutes but the other does not.

Example. Let

$$A = \begin{bmatrix} 0 & \sqrt{2} \\ \sqrt{2} & 0 \end{bmatrix}.$$

If we take $S \equiv A$ and $Q \equiv I$, then $A = QS$ and the factors commute. However,

$$A = \begin{bmatrix} 0 & \sqrt{2} \\ \sqrt{2} & 0 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \equiv \hat{Q}\hat{S},$$

but

$$\hat{S}\hat{Q} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{2} \end{bmatrix} \neq A.$$

The following theorem gives a set of conditions that are sufficient to guarantee that some factorization $A = QS$ of a given matrix $A \in M_n$ has commuting factors.

THEOREM 3. *Let $A \in M_n$ be such that $A^TA = AA^T$. If* rank $A =$ rank A^TA *and A^TA is diagonalizable, then there exists a complex orthogonal $Q \in M_n$ and a symmetric $S \in M_n$ such that $A = QS$ and Q commutes with S.*

Proof. Suppose $A \in M_n$ satisfies all the hypotheses of the theorem. By [2, Cor. 4] there exists an orthogonal $Q \in M_n$ such that

$$A = Q \begin{bmatrix} \lambda_1 P_1 & & & 0 \\ & \lambda_2 P_2 & & \\ & & \ddots & \\ 0 & & & \lambda_k P_k \end{bmatrix} Q^T$$

where $P_i \in M_{n_i}$ is orthogonal and $\lambda_i \in \mathbb{C}$ for $i = 1, 2, \cdots, k$, and $n_1 + n_2 + \cdots + n_k = n$. We can write $A = Q\Lambda PQ^T$ where

$$\Lambda = \begin{bmatrix} \lambda_1 I_1 & & & 0 \\ & \lambda_2 I_2 & & \\ & & \ddots & \\ 0 & & & \lambda_k I_k \end{bmatrix}, \qquad I_i \in M_{n_i}$$

is diagonal and

$$P = \begin{bmatrix} P_1 & & & 0 \\ & P_2 & & \\ & & \ddots & \\ 0 & & & P_k \end{bmatrix}$$

is orthogonal. Then $A = Q(PQ^T)(PQ^T)^T\Lambda PQ^T = Q_1S$ where $Q_1 = QPQ^T$ is orthogonal and $S = (PQ^T)^T\Lambda(PQ^T)$ is symmetric. Now compute

$$\begin{aligned} SQ_1 &= (PQ^T)^T\Lambda(PQ^T)QPQ^T \\ &= QP^T\Lambda PPQ^T \\ &= QP^TP\Lambda PQ^T \quad \text{(since } \Lambda \text{ commutes with } P\text{)} \\ &= Q\Lambda PQ^T = A = Q_1S. \end{aligned}$$ □

The rank and diagonalizability conditions of the theorem, although sufficient, are not necessary. Any symmetric matrix that fails to satisfy either or both of these two hypotheses has a trivial QS factorization in which the factors commute; take the matrix itself as the symmetric matrix and the identity matrix as the orthogonal matrix. For example,

$$A = \begin{bmatrix} 1+i & 1 \\ 1 & 1-i \end{bmatrix}$$

is nonsingular but $AA^T = A^2$ is not diagonalizable, and

$$A = \begin{bmatrix} i & 1 \\ 1 & -i \end{bmatrix}$$

has $AA^T = A^2 = 0$ diagonalizable but fails the rank condition. Moreover, it is known [4, Vol. II, Thm. 3 of Chap. XI] that any nonsingular $A \in M_n$ (which therefore satisfies rank A = rank A^TA) can be written in at least one way as $A = QS$, in which the factors commute if and only if $A^TA = AA^T$, whether or not A^TA is diagonalizable. The rank condition is equivalent to the assumption that the column space of A is nonsingular, i.e., has a rectanormal basis [1], [8, Prop. 157.1].

The first hypothesis of Theorem 2 is that the symmetric matrix A^TA has a symmetric square root, but it is not necessary to assume both that (a) A^TA has a (not necessarily unique) square root and (b) among its square roots there is one that is symmetric. The following result shows that the second statement follows from the first, and leads to a simple sufficient condition for the existence of a QS factorization.

THEOREM 4. *If a symmetric matrix $S \in M_n$ has a square root $B \in M_n$, then it has a symmetric square root that is similar to B.*

Proof. Let $S \in M_n$ be symmetric and suppose $B \in M_n$ satisfies $B^2 = S$. Because every square matrix is similar to a symmetric matrix [5, Thm. 4.4.9], there is a nonsingular $R \in M_n$ such that $B = R\hat{S}R^{-1}$ and $\hat{S} = \hat{S}^T$. Therefore, $S = B^2 = R\hat{S}R^{-1}R\hat{S}R^{-1} = R\hat{S}^2R^{-1}$. Since S and $\hat{S}^2$ are similar and both are symmetric, there is a complex orthogonal $Q \in M_n$ such that $S = Q\hat{S}^2Q^T = Q\hat{S}Q^TQ\hat{S}Q^T = \hat{B}^2$ [1, Thm. 3.6], [4, Vol. II, Chap. XI], [6, Chap. 6]. Thus, $\hat{B} \equiv Q\hat{S}Q^T$ is a symmetric square root of S and $\hat{B}$ is orthogonally similar to $\hat{S}$, which is similar to B. □

The Jordan canonical form of a matrix $A \in M_n$ helps one to determine whether A has a square root or not. If A is nonsingular, it always has a square root. If A is singular, let $A \equiv RJR^{-1}$, where $J \in M_n$ is the Jordan canonical form of A. Let $J_{n_1}(0) \oplus J_{n_2}(0) \oplus \cdots \oplus J_{n_k}(0)$ be the singular part of J, where each summand is a nilpotent Jordan block

$$J_m(0) \equiv \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & 0 \\ & & \ddots & \ddots & \\ & 0 & & 0 & 1 \\ & & & & 0 \end{bmatrix} \in M_m,$$

and let $n_1 \geqq n_2 \geqq \cdots \geqq n_k$. Define $\Delta_1 = n_1 - n_2$, $\Delta_3 = n_3 - n_4, \cdots$. The A has a square root if and only if all $\Delta_i = 0$ or 1 for $i = 1, 3, 5, \cdots$. If k is odd, then we must also require that $n_k = 1$ [3]. See also [6]. Since the numbers $n_1, n_2, \cdots, n_k$ can be determined easily from the numbers rank A^i, $i = 1, 2, \cdots, n$, it is in principle easy to determine whether a given matrix $A \in M_n$ has a square root or not. If a given symmetric matrix A has a square root B with a given rank, then the preceding theorem ensures that A has a symmetric square root with the same rank as B.

We have already observed that if a given matrix $A \in M_n$ can be written as $A = QS$, then A^TA has a square root (namely, S) that has the same rank as A. This necessary condition is not sufficient, however, as the example

$$A = \begin{bmatrix} 1 & 0 \\ i & 0 \end{bmatrix}$$

illustrates. We have

$$A^TA = 0 = B^2 \quad \text{where } B = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}$$

has the same rank as A. Nevertheless, this matrix A cannot be written as $A = QS$ because $A^TA = 0$ is not similar to

$$AA^T = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}.$$

With one additional hypothesis, however, the square root condition is sufficient to guarantee that $A = QS$.

THEOREM 5. *Let $A \in M_n$ be given. If* rank A = rank A^TA *and if A^TA has a square root with the same rank as A, then there exists an orthogonal matrix Q and a symmetric matrix S such that $A = QS$. Conversely, if $A = QS$ with Q orthogonal and S symmetric then A^TA has a square root* (*namely, S*) *with the same rank as A.*

Proof. Let rank $A = k =$ rank A^TA. Using the arguments preceding Theorem 4, we may assume without loss of generality that the first k columns of A are linearly independent. Write $A = [\hat{A} \mid \hat{A}C]$, where $\hat{A} \in M_{n,k}$ has full rank and $C \in M_{k,n-k}$. If A^TA has a square root B with rank B = rank A, we know from Theorem 4 that A^TA has a symmetric square root S with rank S = rank B = rank A. Let $S = [\hat{S} \mid \tilde{S}]$, where $\hat{S} \in M_{n,k}$. Calculate

$$A^TA = \left[\begin{array}{c} \hat{A}^T \\ \hdashline C^T\hat{A}^T \end{array}\right] [\hat{A} \mid \hat{A}C] = \left[\begin{array}{c:c} \hat{A}^T\hat{A} & \hat{A}^T\hat{A}C \\ \hdashline C^T\hat{A}^T\hat{A} & C^T\hat{A}^T\hat{A}C \end{array}\right] = [E \mid EC]$$

where

$$E \equiv \left[\begin{array}{c} \hat{A}^T\hat{A} \\ \hdashline D^T\hat{A}^T\hat{A} \end{array}\right] \in M_{n,k}$$

has full rank. Also $S^2 = A^TA$, i.e., $S[\hat{S} \mid \tilde{S}] = [S\hat{S} \mid S\tilde{S}] = [E \mid EC]$. Therefore, $S\hat{S} = E$ and $S\tilde{S} = EC$. Since $S\hat{S} = E$ has full rank, $\hat{S}$ has full rank k (which is also the rank of S) and hence there exists $F \in M_{k,n-k}$ such that $\tilde{S} = \hat{S}F$. But then $S\hat{S}F = S\tilde{S} = EC = S\hat{S}C$, and hence $S\hat{S}(F - C) = 0$, or $E(F - C) = 0$. Because E has full rank, we have $F - C = 0$, or $F = C$. Thus, $\tilde{S} = \hat{S}C$. By Theorem 2 there exists an orthogonal matrix Q such that $A = QS$. The converse is immediately apparent. □

The following examples show that the two hypotheses of the preceding theorem are independent.

Example. Let

$$A = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}.$$

Since A is symmetric, it can always be expressed as $A = QS$ and A itself is a square root of A^TA. But rank $A \neq$ rank A^TA since $A^TA = 0$.

Example. Let

$$A = \begin{bmatrix} 1 & i \\ 0 & 0 \end{bmatrix}.$$

Then

$$A^TA = \begin{bmatrix} 1 & 0 \\ i & 0 \end{bmatrix} \begin{bmatrix} 1 & i \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix},$$

and so rank A = rank A^TA. However, A^TA does not have a square root because the Jordan form of A^TA is

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

As a practical matter, our results give only a partial solution to the problem of characterizing those singular $A \in M_n$ such that $A = QS$. If A is such that A^TA has a square root at all (easily checked), then we know it has a symmetric square root S. However, Theorem 2 does not offer any way to use the given data (the matrix A) to determine whether this symmetric matrix S, for which we have no explicit representation, satisfies the second condition (2) or (2′) of Theorem 2. Since $A^TA = SQ^TQS = S^2$ is similar to $AA^T = QS^2Q^T$ if $A = QS$, and since two symmetric matrices are similar if and only if they are orthogonally similar, it is tempting to conjecture that *$A = QS$ if and only if A^TA has a square root and A^TA is similar to AA^T.* Since we have neither a proof nor a counterexample to offer, we leave this conjecture as an open problem.

Every square complex matrix has both a polar decomposition and a singular value decomposition. It is possible, however, for a square complex matrix to have a factorization of the form $A = QS$ but not one of the form $A = P_1\Lambda P_2^T$, where P_1 and P_2 are orthogonal and Λ is diagonal. For example the symmetric matrix

$$A = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}$$

has the trivial QS factorization $A = IA \equiv QS$. If $A = P_1\Lambda P_2^T$, then $AA^T = P_1\Lambda^2P_1^T = 0$ implies $\Lambda^2 = 0 = \Lambda$, which implies $A = 0$. Thus, this matrix has a QS factorization but not a factorization analogous to the singular value decomposition. There is no example of the converse phenomenon, for if $A = P_1\Lambda P_2^T$, then $A = (P_1P_2^T)(P_2\Lambda P_2^T) = QS$. See [2] for more results about factorizations of the form $P_1\Lambda P_2^T$.

REFERENCES

[1] D. Choudhury and R. A. Horn, *An analog of the Gram-Schmidt algorithm for complex bilinear forms and diagonalization of complex symmetric matrices.* Technical Report No. 454, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, January 2, 1986.

[2] ———, *An analog of the singular value decomposition for complex orthogonal equivalence*, Linear and Multilinear Algebra, to appear.

[3] G. W. Cross and P. Lancaster, *Square roots of complex matrices*, Linear and Multilinear Algebra, 1 (1974), pp. 289–293.

[4] F. R. Gantmacher, *The Theory of Matrices*, Vols. I and II, Chelsea, New York, 1977.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, New York, 1985.

[6] ———, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, New York, 1987.

[7] O. T. O'Meara, *Introduction to Quadratic Forms*, Springer–Verlag, Berlin, New York, 1973.

[8] E. Snapper and R. J. Troyer, *Metric Affine Geometry*, Academic Press, New York, 1971.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

009

FURTHER COMPARISONS OF DIRECT METHODS FOR COMPUTING STATIONARY DISTRIBUTIONS OF MARKOV CHAINS*

DANIEL P. HEYMAN†

Abstract. An algorithm for computing the stationary distribution of an irreducible Markov chain consisting of ergodic states is described in Grassmann et al. [Oper. Res., 33 (1985), pp. 1107–1116]. In this algorithm, all the arithmetic operations use only nonnegative numbers and there are no subtractions. In this paper we present numerical evidence to show that this algorithm achieves significantly greater accuracy than other algorithms described in the literature. We also describe our computational experience with large block-tridiagonal matrices.

Key words. direct methods, Markov chain, sparsity schemes, Gaussian elimination, block-tridiagonal matrices

AMS(MOS) subject classifications. 65U05, 60-04, 60J10, 60J27

Introduction. The solution of many probabilistic models requires the computation of the stationary distribution of a finite-state Markov chain. A direct (i.e., not iterative) algorithm to compute the stationary distribution of a finite-state Markov chain consisting of ergodic states is given in Grassmann, Taksar and Heyman [2]. For brevity, we will call this the GTH algorithm. This paper describes the numerical accuracy of the GTH algorithm, and compares its performance to several alternatives described in Harrod and Plemmons [3]. For the test problems considered by Harrod and Plemmons, the GTH algorithm provides greater accuracy than the alternatives described by those authors.

Algorithms are often evaluated by their speed and their storage requirements. The GTH algorithm requires about $2N^3/3$ operations ($N + 1$ is the number of states) and stores the transition matrix and the vector of stationary probabilities. Large problems frequently possess a special structure that can be exploited to reduce execution time and storage requirements. The block-tridiagonal (or generalized birth-and-death) structure occurs in many queueing models when the state space has two dimensions. We describe how to modify the GTH algorithm to exploit this structure. An example with 3,600 states requires approximately 7 minutes of execution time on a VAX 8600, so large problems of this type can be solved.

This note contains three sections. Section 1 is a review of the GTH algorithm. Section 2 compares the accuracy of the GTH algorithm to the algorithms considered by Harrod and Plemmons. Section 3 describes some computational experiences with large tri-diagonal transition matrices.

1. Review of the GTH algorithm. Let P be the transition matrix of a Markov chain with states 0, 1, $\cdots$, N, and let π be a stationary distribution of P, i.e.,

$$\pi = \pi P \quad \text{and} \quad \sum \pi_i = 1.$$

The GTH algorithm for obtaining π is as follows:

1. For $n = N, N - 1, \cdots, 1$, do the following:

* Received by the editors January 20, 1986; accepted for publication (in revised form) August 21, 1986.

† Bell Communications Research, Red Bank, New Jersey 07701.

$$\text{Let } S = \sum_{j=0}^{n-1} p_{nj}.$$

$$\text{Let } p_{in} = p_{in}/S, \qquad i < n.$$

$$\text{Let } p_{ij} = p_{ij} + p_{in}p_{nj}, \qquad i,j < n.$$

2. Let TOT $= 1$ and $\pi_0 = 1$.
3. For $j = 1, 2, \cdots, N$ do the following:

$$\text{Let } \pi_j = p_{0j} + \sum_{k=1}^{j-1} \pi_k p_{kj}.$$

$$\text{Let TOT} = \text{TOT} + \pi_j.$$

4. Let $\pi_j = \pi_j/\text{TOT}, j = 0, 1, \cdots, N$.

Notice that all of the arithmetic operations use only nonnegative numbers, and that there are no subtractions. Grassmann [1] shows that algorithms with this property are extremely resistant to rounding errors. The alternative methods do not have the above property, and only 1 or 2 decimal digits are accurately obtained in a chain with merely 5 states (see Test Problem 3 in § 2). The GTH algorithm is accurate to 7 decimal digits in this example.

Grassmann et al. show that this algorithm produces the limiting distribution when P is irreducible and consists of aperiodic positive-recurrent states. However, the algorithm will work when the states are periodic, in which case it produces the stationary (but not limiting) distribution. If P has transient states in addition to an irreducible set of positive-recurrent states, the algorithm may fail because step 1 produces $S = 0$. This will not occur if the states are renumbered so that the transient states have the largest numbers.

Grassmann et al. observe that the algorithm can be applied to a continuous-time Markov chain (CTMC) by appealing to the uniformization procedure. For a CTMC with generator (i.e., rate matrix) Q, one applies the algorithm to the matrix that agrees with Q on the off-diagonal elements and has zeros on the diagonal.

2. Accuracy of the GTH algorithm. Harrod and Plemmons compare three direct methods. Two of them are the recommended choices from the comparisons in Paige, Styan and Wachter [7]; the third is the authors' partition factorization method. The latter method requires about one-half the number of operations compared to the GTH method. They consider 5 test problems. The fifth one is not completely specified in the paper and is no longer available, so we will use a comparable problem.

Harrod and Plemmons assess the accuracy of an algorithm by first computing π in double-precision arithmetic using the QR algorithm as implemented in the LINPACK software package. This is considered to be the "true" value. (The QR algorithm is quite computationally bound, requiring about N^3 operations.) Then $\hat{\pi}$ is computed in single-precision arithmetic. The relative error is $\sum|\pi_i - \hat{\pi}_i|$. For each test problem, we will let MinRE be the minimum relative error for the methods considered by Harrod and Plemmons, and let MaxRE be the maximum relative error produced by that method. (Since one of the equations in $\pi = \pi P$ is redundant, the performance may depend on which equation is dropped.) The relative error of the GTH algorithm will be denoted by GTHRE.

Test Problem 1. The transition matrix is:

$$\begin{bmatrix} .2 & 0 & 0 & .6 & 0 & 0 & 0 & 0 & 0 & .2 \\ 0 & .1 & 0 & 0 & .6 & 0 & .3 & 0 & 0 & 0 \\ 0 & .1 & 0 & 0 & 0 & 0 & 0 & .8 & 0 & .1 \\ 0 & 0 & .6 & 0 & .3 & 0 & 0 & 0 & 0 & .1 \\ 0 & .5 & 0 & 0 & .5 & 0 & 0 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 & .2 & 0 & 0 & 0 & .3 & 0 \\ 0 & 0 & 0 & 0 & .7 & 0 & .2 & 0 & 0 & .1 \\ .1 & 0 & .9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .1 & 0 & 0 & 0 & .8 & 0 & 0 & 0 & .1 \\ 0 & .4 & 0 & 0 & 0 & .4 & 0 & 0 & 0 & .2 \end{bmatrix}.$$

The GTH algorithm determines that states 0, 2, 3, and 7 are transient. (Notice that step 1 produces $S = 0$ when $n = 1$.) The GTH algorithm produces 6 significant decimal digits while some of the alternatives produce only 5. We have

GTHRE	MinRE	MaxRE
4.5×10^{-8}	6.9×10^{-8}	3.7×10^{-6}

and the GTH algorithm provides greater accuracy than the alternatives.

Test Problem 2. The transition matrix is

$$\begin{bmatrix} .85000 & 0 & .14900 & .00090 & 0 & .00005 & 0 & .00005 \\ .10000 & .65000 & .24900 & 0 & .00090 & .00005 & 0 & .00005 \\ .10000 & .80000 & .09960 & .00030 & 0 & 0 & .00010 & 0 \\ 0 & .00040 & 0 & .70000 & .29950 & 0 & .00010 & 0 \\ .00050 & 0 & .00040 & .39900 & .60000 & .00010 & 0 & 0 \\ 0 & .00005 & 0 & 0 & .00005 & .60000 & .24990 & .15000 \\ .00003 & 0 & .00003 & .00004 & 0 & .10000 & .80000 & .09990 \\ 0 & .00005 & 0 & 0 & .00005 & .19990 & .25000 & .55000 \end{bmatrix}.$$

(We have corrected a typographical error on element (1, 5).) The GTH algorithm produces 6 significant decimal digits while the alternatives produce between 3 and 6. We have

GTHRE	MinRE	MaxRE
9.64×10^{-8}	3.96×10^{-5}	1.71×10^{-4}

and the GTH algorithm provides greater accuracy than the alternatives.

This matrix is also considered in Koury, McAllister and Stewart [6]. For the approximate method they recommend, (VANTD), an estimated 173 floating point operations are required to obtain a solution. The GTH algorithm requires 400 floating point operations.

Test Problem 3. The transition matrix is

$$\begin{bmatrix} .999999 & 1.0\times 10^{-7} & 2.0\times 10^{-7} & 3.0\times 10^{-7} & 4.0\times 10^{-7} \\ .4 & .3 & 0 & 0 & .3 \\ 5.0\times 10^{-7} & 0 & .999999 & 0 & 5.0\times 10^{-7} \\ 5.0\times 10^{-7} & 0 & 0 & .999999 & 5.0\times 10^{-7} \\ 2.0\times 10^{-7} & 3.0\times 10^{-7} & 1.0\times 10^{-7} & 4.0\times 10^{-7} & .999999 \end{bmatrix}.$$

The GTH algorithm produces 6 significant decimal digits while the alternatives produce only 1 or 2. We have

GTHRE	MinRE	MaxRE
3.1×10^{-8}	6.12×10^{-3}	3.89×10^{-2}

and the GTH algorithm provides greater accuracy than the alternatives.

Test Problem 4. The transition matrix is

$$\begin{bmatrix}
.1-\varepsilon & .3 & .1 & .2 & .3 & \varepsilon & 0 & 0 & 0 & 0 \\
.2 & .1 & .1 & .2 & .4 & 0 & 0 & 0 & 0 & 0 \\
.1 & .2 & .2 & .4 & .1 & 0 & 0 & 0 & 0 & 0 \\
.4 & .2 & .1 & .2 & .1 & 0 & 0 & 0 & 0 & 0 \\
.6 & .3 & 0 & 0 & .1 & 0 & 0 & 0 & 0 & 0 \\
\varepsilon & 0 & 0 & 0 & 0 & .1-\varepsilon & .2 & .2 & .4 & .1 \\
0 & 0 & 0 & 0 & 0 & .2 & .2 & .1 & .3 & .2 \\
0 & 0 & 0 & 0 & 0 & .1 & .5 & 0 & .2 & .2 \\
0 & 0 & 0 & 0 & 0 & .5 & .2 & .1 & 0 & .2 \\
0 & 0 & 0 & 0 & 0 & .1 & .2 & .2 & .3 & .2
\end{bmatrix}.$$

(We have written the $-\varepsilon$ terms that failed to appear in the paper by Harrod and Plemmons.) This matrix is reducible when $\varepsilon = 0$, so it potentially might cause the GTH algorithm difficulty. The GTH algorithm produced 6 decimal digits accurately. (The corresponding figures for the alternatives are not reported.) We have

ε	GTHRE	MinRE	MaxRE
10^{-1}	1.38×10^{-7}	1.18×10^{-6}	6.74×10^{-6}
10^{-3}	1.38×10^{-7}	6.16×10^{-5}	5.51×10^{-4}
10^{-5}	1.38×10^{-7}	7.03×10^{-3}	3.90×10^{-2}
10^{-7}	1.38×10^{-7}	6.42×10^{-1}	8.29×10^{-1}

A novel feature of this matrix is that the GTH algorithm produces the same answer for each of the four values of ε given above, in both single and double precision arithmetic. That property persists in double precision (but not quite in single precision) as the states are renumbered. The explanation is that π is independent of ε, as we will now show.

Since an ε appears in only two columns of the transition matrix, only two balance equations contain an ε. They are

(1) $$.9\pi_0 = .2\pi_1 + .1\pi_2 + .4\pi_3 + .6\pi_4 + \varepsilon(\pi_5 - \pi_0)$$

and

(2) $$.9\pi_5 = .2\pi_6 + .1\pi_7 + .5\pi_8 + .1\pi_9 - \varepsilon(\pi_5 - \pi_0).$$

In the steady state, we must have (see, e.g., Theorem 7-13 in Heyman and Sobel [4])

$$\sum_{i=0}^{4} \pi_i \sum_{j=5}^{9} p_{ij} = \sum_{i=5}^{9} \pi_i \sum_{j=0}^{4} p_{ij}$$

which is conservation of flow between the sets of states $\{0, 1, 2, 3, 4\}$ and $\{5, 6, 7, 8, 9\}$. In this chain, the equation above is simply

(3) $$\pi_0\varepsilon = \pi_5\varepsilon$$

which shows that $\pi_0 = \pi_5$ for every $\varepsilon > 0$. Substituting (3) into (1) and (2) shows that (1) and (2) are independent of ε, and hence π is independent of ε.

Test Problem 5. Test problem 5 of Harrod and Plemmons is an 84-state chain used in a job line production model. The positions of the nonzero elements is determined by the model, and their values are randomly generated. The values they used are no longer available, so I used a Markov chain that arose in a communications network model I was working on.

For 76 states, the positions of the positive elements of the transition matrix are shown in Fig. 1. All of the elements of π are correct to 5 decimal digits. For the

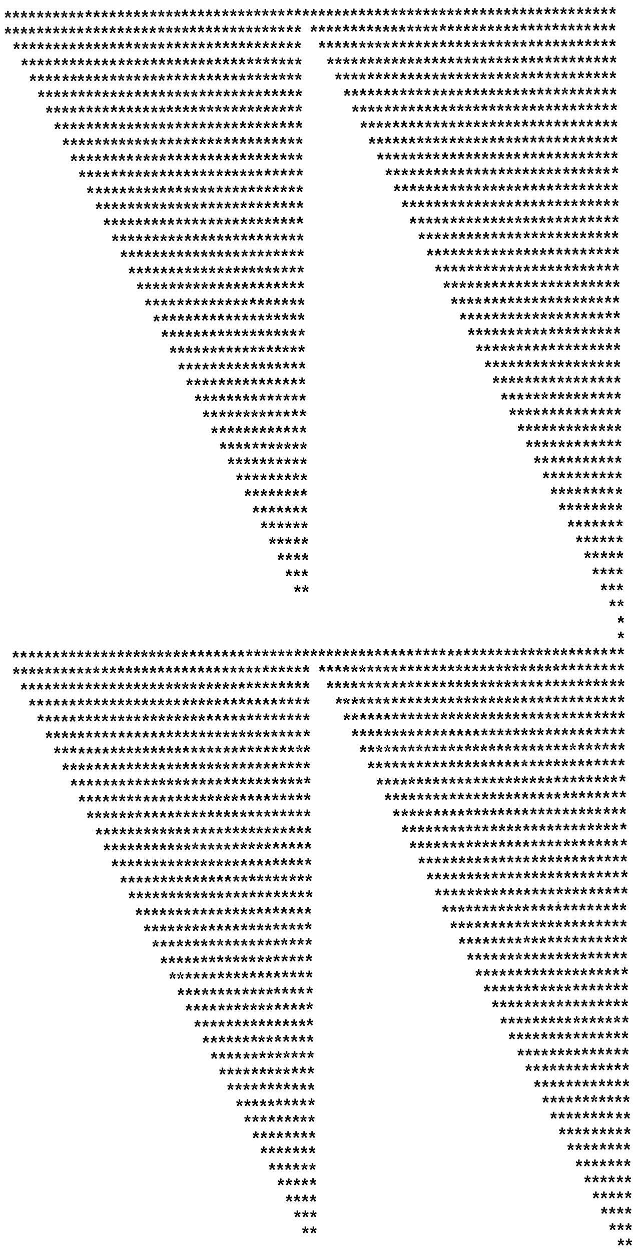


FIG. 1. *Nonzero elements of P.*

same model with 102 states, 5 digit accuracy is obtained whenever $\pi_i \geqq 10^{-33}$. When $\pi_i < 10^{-33}$, there may be no accurate decimal digits (e.g., $\pi_{45} = 0.622263 \times 10^{-38}$ in double precision and is 0.000000 in single precision) because single precision cannot represent such a small number. In most situations, these discrepancies are inconsequential.

3. Computing with large block tri-diagonal matrices. Many congestion models, especially systems of overflowing queues, have a transition matrix that is block tri-diagonal, e.g.,

$$P=\begin{bmatrix} B_0 & D_0 & & & & \\ C_1 & B_1 & D_1 & & & \\ & C_2 & B_2 & D_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & C_{K-1} & B_{K-1} & D_{K-1} \\ & & & & C_K & B_K \end{bmatrix}.$$

The blocks B_k, D_k and C_k are square and they all have the same number of elements. Frequently, each C_k is upper-triangular and each D_k is lower triangular. We will only consider this situation. Examples are given in Kaufman [5] and Sumita and Shanthikumar [8]. In both of these examples there are two queues. The subscript on each block is the level of one queue and the components of the block show the transition probabilities for the other queue. Kaufman also shows that when a third queue is added to the model, the block tri-diagonal structure remains.

It is clear that this structure should be exploited when attempting to solve chains with many states. We will now see how the GTH algorithm exploits this structure and some large problems that have been solved will be described.

When there are positive integers g and h such that $p_{ij} = 0$ for $j < i - g$ and $j > i + h$ for every i, the transition matrix is called *banded.* Grassmann et al. [2] show that the GTH algorithm preserves this property and that steps 1 and 3 of the algorithm can be streamlined. Notice that block tridiagonal matrices are necessarily banded, so the comments above apply to them. Moreover, every banded matrix can be written as a block tri-diagonal matrix with the C's upper-triangular and the D's lower triangular.

When each block is $s \times s$, the storage requirements for the matrix are $(3K + 1)s^2$ numbers for the special structure compared to n^2 for the general case, where $n = s(K + 1)$. For example, when $s = K + 1 = 50$ and $n = 2{,}500$, the special structure stores 370,000 numbers, which is 6% of the 6,250,000 numbers stored without exploiting the structure.

For the special structure, the three parts of step 1 of the algorithm take s additions, followed by s multiplications and then an addition and a multiplication are done s^2 times, for a total of $2s(s + 1)$ operations. This is done $n - 1$ times. Step 3 requires $2s$ multiplications and additions followed by an addition. This is done $n - 1$ times. Step 4 contains one multiplication and is done n times. Thus, there are about (actually, slightly less than) $2s(s + 1)^2(K + 1)$ operations. When $s = K + 1 = 50$, the special structure needs about 13 million operations compared to roughly 10 trillion operations if the structure is not exploited.

For $n = s(K + 1)$ held fixed, both storage and computing time decrease as s decreases.

Some examples. The transition matrix I used is the jump chain derived from the generator matrix (2.3) in Kaufman [5]. The model consists of two queues, each with 5 servers and b waiting positions. A customer that tries to enter queue 1 when all of its waiting positions are occupied will attempt to join the second queue. Let λ_i be the arrival rate (including lost customers but not including overflow customers) and μ_i be the service rate at queue i. The blocks are

$$C_k = \mu_i \min(k, 5)I, \qquad k = 1, 2, \cdots, b,$$

$$D_k = \lambda_1 I, \qquad k = 0, 1, \cdots, b-1,$$

where I is the $(b + 1) \times (b + 1)$ identity matrix. For $k = 0, 1, \cdots, b - 1$, and $i = 1, 2, \cdots, b$,

$$B_k(i, i-1) = \mu_2 \min(i, 5), \qquad B_k(i-1, i) = \lambda_2,$$

and is 0 otherwise. Finally,

$$B_b(i, i-1) = \mu_2 \min(i, 5), \qquad B_b(i-1, i) = \lambda_1 + \lambda_2,$$

and is 0 otherwise. Here, $s = K + 1 = b + 1$ and $n = (b + 1)^2$.

The general purpose GTH algorithm and a special purpose version that exploits the special structure were programmed in FORTRAN 77 and run on 3 different computers. Various values of b were used. The running times reported below include the time to create the matrices.

When $b = 9$, the running times for the special purpose algorithm were

VAX 11/780	PYRAMID	VAX 8600
2.4 sec	1.9 sec	0.3 sec .

The running times for the general purpose algorithm were

VAX 11/780	PYRAMID	VAX 8600
14.5 sec	15.9 sec[1]	2.9 sec .

The remaining times are for the special purpose algorithm only:

	VAX 11/780	PYRAMID	VAX 8600
$n = 2{,}500$	10:18.4	6:10.9	3:11.5
$n = 3{,}600$	22:39.5	24:42.6[1]	6:45.58
$n = 4{,}900$	-	1 hr 9:53.3	- .

Only the Pyramid, which has virtual memory, was capable of running the $n = 4{,}900$ example.

The limited experience reported here suggests that the special purpose algorithm is capable of handling models within the three thousand state range.

The very special structure of the blocks (C and D are diagonal and B is tridiagonal) can be used to obtain greater efficiency. I have not done so because it is the tridiagonal structure of P that was of interest. The very special structure should not have affected the computation times for the algorithm that was used.

REFERENCES

[1] WINFRIED K. GRASSMANN, *Rounding errors in some recursive methods used in computational probability*, Tech. Rpt. 73, Dept. of Operations Research, Stanford University, 1983.

[2] WINFRIED K. GRASSMANN, MICHAEL I. TAKSAR AND DANIEL P. HEYMAN, *Regenerative analysis and steady state distributions for Markov chains*, Oper. Res., 33 (1985), pp. 1107–1116.

[3] W. J. HARROD AND R. J. PLEMMONS, *Comparisons of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 453–469.

[4] DANIEL P. HEYMAN AND MATTHEW J. SOBEL, *Stochastic Models in Operations Research*, Vol. I, McGraw-Hill, New York, 1982.

[5] LINDA KAUFMAN, *Matrix methods for queueing problems*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 525–552.

[6] J. R. KOURY, D. F. MCALLISTER AND W. J. STEWART, *Iterative methods for computing stationary distributions of nearly completely decomposable Markov chains*, this Journal, 5 (1984), pp. 164–186.

[7] C. C. PAIGE, P. H. STYAN AND P. G. WACHTER, *Computation of the stationary distribution of a Markov chain*, J. Statist. Comput. Simulation, 4 (1975), pp. 173–186.

[8] U. SUMITA AND J. G. SHANTHIKUMAR, *APL software development for central telephone switching systems via the row-continuous Markov chain procedure*, Proc. Eleventh International Teletraffic Congress, Kyoto, 1985.

[1] The floating point hardware on the Pyramid changed during the course of these experiments, and that apparently caused it to become slower.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

SOLUTION OF A LINEAR RECURRENCE EQUATION ARISING IN THE ANALYSIS OF SOME ALGORITHMS*

WOJCIECH SZPANKOWSKI†

Abstract. We study a recurrence equation of type

$$l_n(2^{n+s}-2)=2^n a_n+\sum_{k=1}^{n-1}\genfrac{}{}{0pt}{}{n}{k}\, l_k,\qquad n\geqq N$$

where a_n is any sequence and s, N are integers. This type of recurrence arises in many applications in computer sciences and telecommunications, e.g., in the analysis of unsuccessful search in a family of Patricia trees, in the average complexity of an algorithm generating exponentially distributed variates, in trie statistics, in the performance evaluation of conflict resolution algorithms in a broadcast communication environment, etc. We present a closed-form solution of the recurrence and then we establish an asymptotic approximation for it. In addition, we offer an approximation of a generating function, $l(z)$, of l_n for small values of z.

Key words. linear recurrence, Patricia trees, conflict resolution algorithms, Bernoulli numbers, Bernoulli polynomials, Bernoulli inverse relationship, asymptotic approximation, Mellin transform

AMS(MOS) subject classifications. 11B68, 39A10, 39B20, 41A60, 68Q25, 68R99

1. Introduction. Let the infinite sequence $l_0, l_1, l_2, \cdots$, satisfy the following linear recurrence

$$f_n l_n=a_n+\gamma\sum_{k=0}^{n-1}\genfrac{}{}{0pt}{}{n}{k}\, p^k q^{n-k} l_k,\qquad p+q=1,\tag{1.1}$$

where a_n is a given sequence, and γ is a constant. The coefficient f_n is either $1-\gamma p^n$ or $1-\gamma p^n-\gamma q^n$. Such recurrence arises quite often in practice [2], [4], [6]–[8], [10], [12], [15], [16], [18] with the *additive term*, a_n, in (1.1) appropriately chosen.

Let first $f_n = 1-\gamma p^n$, and $\gamma = 1$. Consider a digital search trie [12] with n records (external nodes). Then l_n represents the external path length in such a tree if $a_n = n$, and depth of a leaf if $a_n = 1$. An annoying flaw of the digital search trie is "one-way branching," which leads to the creation of extra nodes in the tree. To avoid this, D. R. Morrison discovered a data structure which he named the Patricia tree [12]. The external path length in the Patricia is also given by (1.1) with $a_n = n(1-p^n-q^n)$. Other applications of (1.1) have been found recently in the telecommunications field where the so-called *conflict resolution algorithms* for broadcast communication are studied [4], [10], [15]. Here, either $a_n = 1$ or $a_n = 1-p^n$ [10], [15]. More examples can be found in [8] and [12]. Note that all of these cases may be treated in a unified manner if one solves (1.1) with the additive term a_n being any sequence of numbers. This was done in [15] where the general solution of (1.1) with an asymptotic approximation is presented.

In this paper we study (1.1) with $f_n = 1-\gamma p^n-\gamma q^n$. This is almost identical to the one described above, but the appearance of γq^n is enough to change the entire character of the recurrence, and the methods used before are wiped out. Because of that we are forced to further restrict the class of (1.1). We assume throughout the paper that $p = q = 0.5$, but we present the solution of (1.1) for any sequence of a_n. Such a restriction does not limit the applications of (1.1). For example, it turns out that an unsuccessful search

* Received by the editors July 8, 1985; accepted for publication (in revised form) August 21, 1986.

† Department of Computer Science, Purdue University, West Lafayette, Indiana 47907, and Technical University of Gdansk, Gdansk, Poland.

in the Patricia tree is described by (1.1) with $\gamma = 1$. For the average value of the unsuccessful search we must assume $a_n = 1 - 2^{1-n}$ [12], and for higher moments the additive term is more complex, but still (1.1) is to be considered. Flajolet and Saheb [6] studied (1.1) with $\gamma = 0.5$ and some particular form of a_n, analyzing the complexity of generating an exponentially distributed variate. Recently, (1.1) has found applications in the performance evaluation of the Gallager–Tsybakov–Mikhailov conflict resolution algorithm [6], [18]. In this case either $a_n = 1$, $\gamma = 1$ or $a_n = 2^{-n-1}$ and $\gamma = 0.5$ [16]. The same recurrence might be used to study another conflict resolution algorithm proposed by Berger [2], [16] (for more examples, see [8]).

Under the above assumption we present a closed form solution of (1.1) and asymptotic approximation to it. In addition, an approximation of the exponential generation function of l_n is given. To the author's knowledge such a solution was available only for a few specific values of a_n, namely: Knuth considered $a_n = 1 - 2^{1-n}$, $\gamma = 1$ [12, p. 409] while Szpankowski [16] assumed either $a_n = 1$, $\gamma = 1$ or $a_n = 2^{-n-1}$, $\gamma = 0.5$. This paper generalizes these results.

2. Problem formulation. We shall study (1.1) under the following assumptions:

(a) $p = q = 0.5$,

(b) $f_n = 1 - \gamma p^n - \gamma q^n = 1 - \gamma 2^{-n} - \gamma 2^{-n}$,

(c) *For simplicity* we assume also that $\gamma = 2^{-s}$ where s is an integer (later we shall point out that this assumption is irrelevant for the proposed method).

In addition, instead of a_n we consider $2^n a_n$ for reasons which will be clear later. Then, the problem is to solve the following recurrence:

$$l_0, l_1, \cdots, l_N \textbf{ given}$$

$$(2.1) \qquad (2^{n+s} - 2) l_n = 2^n a_n + \sum_{k=1}^{n-1} \binom{n}{k} l_k, \qquad n > N$$

where s, N are integers such that $N > -s$, and a_n is any sequence.

Example 1a. *Unsuccessful search in a family of Patricia trees* [8], [12]. Digital searching is a well-known technique for storing and retrieving information using lexicographical (digital) structure of words. Let U be an alphabet containing two elements, $U = \{\sigma_1, \sigma_2\}$ and we define a set S which consists of finite numbers, say n, of (possible infinite) strings (keys) from U. A *trie or radix search trie* is a binary digital search tree in which edges are labelled by elements from U and leaves (external nodes) contain the keys [8], [12]. The access path from the root to a leaf is a minimal prefix of the information contained in the leaf. The radix trie has an annoying flaw: there is "one-way branching" which leads to the creation of extra nodes in the tree. D. R. Morrison discovered a way to avoid this problem in a structure which he named the *Patricia tree*. In such a tree all nodes have branching degree equal to two. For more details see [8], [12]. The Patricia tree finds many applications, e.g., in lexicographical order, dynamic hashing algorithms and so on. If we want to store a new element in the tree, then two situations may occur. Either the element is already in the tree, which we call a *successful search*; or the element is not in the tree, hence an *unsuccessful search*. It turns out that the average value of the unsuccessful search, c_n, in a family of Patricia trees with n records satisfies the following recurrence [12, p. 498]:

$$c_0 = c_1 = 0,$$

$$(2.2) \qquad (2^n - 2) c_n = 2^n - 2 + \sum_{k=1}^{n-1} \binom{n}{k} c_k.$$

This is equivalent to (2.1) with $s = 0$, $N = 1$, $a_n = 1 - 2^{1-n}$.

Example 2a. *Gallager–Tsybakov–Mikhailov conflict resolution algorithm* [7], [18]. Assume an infinite number of users sharing a common communication channel. Since the channel is the only means of communication among the users, packet collision is inevitable, if central coordination is not provided. The problem is to find an efficient algorithm for retransmitting conflicting packets. It turns out that the so-called *conflict resolution algorithms* [2], [4], [7] are the most efficient, and among them the Gallager–Tsybakov–Mikhailov (GTM) algorithm [7], [18] achieves the highest throughput. The idea of the algorithm is to partition a conflict of multiplicity n into smaller conflicts by observing the channel and learning whether in the past it was idle, success or collision (ternary feedback). The performance of the algorithm depends on two quantities T_n and W_n, where n is the multiplicity of a conflict. T_n represents the average length of a *conflict resolution interval*, while W_n is the average length of the so-called *resolved interval* (for details see [7], [18]). It is proved that T_n and W_n satisfy recurrences [18]

$$T_0 = T_1 = 1,$$

$$(2^n - 2)T_n = 2^n + nT_{n-1} + \sum_{k=1}^{n-1} \binom{n}{k} T_k \tag{2.3}$$

and

$$W_0 = W_1 = 1,$$

$$(2^{n+1} - 2)W_n = 1 + nW_{n-1} + \sum_{k=1}^{n-1} \binom{n}{k} W_k. \tag{2.4}$$

These recurrences are *not* of type (2.1), but in [16] we have proved that both recurrences might be solved if one finds a solution of the following recurrences

$$t_0 = t_1 = 1,$$

$$(2^n - 2)t_n = 2^n + \sum_{k=1}^{n-1} \binom{n}{k} t_k \tag{2.5}$$

and

$$w_0 = w_1 = 1,$$

$$(2^{n+1} - 2)w_n = 1 + \sum_{k=1}^{n-1} \binom{n}{k} w_k \tag{2.6}$$

which fall into class of (2.1). Moreover, it is shown that for a Poisson message arrival process the maximum throughput $\lambda_{\max}$ of the algorithm is equal to

$$\lambda_{\max} = \max_z \frac{zW(z)}{T(z)} \tag{2.7}$$

where

$$W(z) = \sum_{n=0}^{\infty} W_n \frac{z^n}{n!}, \qquad T(z) = \sum_{n=0}^{\infty} T_n \frac{z^n}{n!}.$$

Note that $W(z)$ and $T(z)$ are exponential generating functions of W_n and T_n, respectively.

Example 3a. *Berger's conflict resolution algorithm* [2]. Consider now a conflict resolution algorithm as in Example 2a with binary feedback, that is, a user distinguishes only two states of a channel: nothing or something. Then Berger [2] described an algorithm for which the average length of the conflict resolution interval T_n and the average length of the resolved interval W_n satisfy recurrences similar to (2.3) and (2.4) except the first term which is either $2^{n+1}+n-1+nT_{n-1}$ or $1+nW_{n-1}$. Neglecting terms nT_{n-1} and nW_{n-1} we have to solve the following equations:

$$t_0=0, \qquad t_1=1,$$

(2.8)
$$(2^n-2)t_n=2^{n+1}+n-1+\sum_{k=1}^{n-1}\binom{n}{k}t_k,$$

$$w_0=w_1=1,$$

(2.9)
$$(2^{n+1}-2)w_n=1+\sum_{k=1}^{n-1}\binom{n}{k}w_k.$$

Moreover, the maximum throughput $\lambda_{\max}$ satisfies

(2.10)
$$\lambda_{\max}=\max\frac{xW(x)}{e^x+T(x)}$$

where $T(z)$ and $W(z)$ are defined as in (2.8) and (2.9), respectively.

These motivating examples suggest that from a practical point of view both a closed form solution of (2.1) and the generating function of the solution are interesting. In particular, for (2.7) and (2.10) it is much more important to derive an approximation of $W(z)$ and $T(z)$ for small values of z, than to obtain exact closed form solution of (2.5) and (2.6).

3. Solution of the recurrence. Let $l(z)$ be the exponential generating function for the sequence l_n, $n=0, 1, \cdots$, defined in (2.1). Let us also introduce a new sequence L_n, $n=0, 1, \cdots$, as follows:

(3.1)
$$L_n=l_n-l_0, \qquad L(z)=l(z)-l_0e^z$$

where $L(z)$ is an exponential generating function of L_n. Note that $L_0=0$, and recurrence (2.1) is transformed into

$$L_0=0, \qquad L_1, \cdots, L_N \text{ given},$$

(3.2)
$$(2^{n+s}-2)L_n=2^na_n+l_0(2^n-3)+\sum_{k=1}^{n-1}\binom{n}{k}L_k, \qquad n>N.$$

To solve (3.2) we use the generating function method. Multiplying (3.2) for $n>N$ by $z^n/n!$ and taking into consideration initial conditions, one shows that

(3.3)
$$\begin{aligned}&2^sL(2z)-L(z)(e^z+1)\\&=a(2z)-a_0-l_0(2^s-1)(e^{2z}-1)\\&+\sum_{k=1}^{N}\frac{z^k}{k!}\left\{L_k(2^{k+s}-1)+l_0(2^s-1)2^k-a_k2^k-\sum_{i=0}^{k}\binom{k}{i}L_i\right\}\end{aligned}$$

where $a(z)$ is the exponential generating function for a_n. Substituting now in (3.3) z by $z/2$ and using (3.1) we obtain

(3.4)
$$L(z)=2^{-s}L(z/2)(e^{z/2}+1)+b(z)$$

where

(3.5a)
$$b(z)=2^{-s}[a(z)-a_0]-l_0(1-2^{-s})(e^z-1)+\sum_{k=1}^{N}\frac{z^k}{k!}g_k,$$

(3.5b)
$$g_k=l_k(1-2^{-k-s})-a_k2^{-s}-2^{-k-s}\sum_{i=1}^{k}\binom{k}{i}l_i, \qquad k=1,2,\cdots,N.$$

To find a closed form solution for (3.2) we must solve the functional equation (3.4). The easiest way is to introduce a new function $H(z)$ as follows:

$$H(z) = L(z)\frac{z}{e^z - 1}. \tag{3.6}$$

Then, (3.4) becomes

$$H(z) = 2^{1-s}H(z/2) + \frac{b(z)z}{e^z - 1}. \tag{3.7}$$

We prove that

LEMMA 3.1. *A general solution of functional equation* (3.7) *is given by*

$$H(z) = H^*(z) + \sum_{k=0}^{\infty} 2^{(1-s)k}\frac{b(z2^{-k})z2^{-k}}{e^{z2^{-k}} - 1} \tag{3.8}$$

where $H^*(z) = \lim_{k\to\infty} 2^{k(1-s)}H(z2^{-k})$, *provided* $H^*(z)$ *exists and the series in* (3.8) *is convergent.*

Proof. Iterating (3.7) n times and taking the limit as $n \to \infty$ we find (3.8), assuming the appropriate limits exist. □

Let us now consider $H^*(z)$. We show the following:

COROLLARY 3.2. *If* $H(z)$ *is differentiable* $(1 - s)^+$ *times at* $z = 0$, *where* $a^+ = \max\{0, a\}$, *then*

$$H^*(z) = z^{(1-s)^+}\frac{H^{(1-s)^+}(0)}{(1-s)^+!}. \tag{3.9}$$

Proof. Assume first $s \geqq 1$ and note that by (3.6) and (3.2) $H(0) = 0$. Then

$$\lim_{k\to\infty} 2^{k(1-s)}H(z2^{-k}) = 0.$$

Now let $s < 1$ and $u = z2^{-k}$. Then applying l'Hospital's rule $1 - s$ times, we find

$$\lim_{k\to\infty} 2^{k(1-s)}H(z2^{-k}) = z^{1-s}\lim_{u\to 0}\frac{H(u)}{u^{1-s}} = z^{1-s}\frac{H^{(1-s)}(0)}{(1-s)!}$$

where $H^{(n)}(z_0)$ is the nth derivative of $H(z)$ at z_0. □

We now present sufficient and necessary conditions for convergence of the series in (3.8). Let b_n, $n = 0, 1, \cdots$, be coefficients in the Taylor expansion of $b(z)$ at $z = 0$. By definition we assume also that $b_{-n} = 0$, $n = 0, 1, \cdots$. Then

COROLLARY 3.3. *The series in* (3.8) *is convergent if and only if*

$$b_0 = b_1 = \cdots = b_{1-s} = 0 \tag{3.10}$$

provided $b(z)$ *is* $(1 - s)^+$*-times differentiable at* $z = 0$.

Proof. Necessity. Let $z > 0$ be a fixed real number and denote the series in (3.8) as $\sum_{k=0}^{\infty} \alpha_k$, which is assumed to be convergent. This implies that $\lim_{k\to\infty} \alpha_k = 0$ [11], i.e., the following must be satisfied:

$$\lim_{k\to\infty} 2^{(1-s)k}\frac{b(z2^{-k})z2^{-k}}{e^{z2^{-k}} - 1} = z^{1-s}\lim_{u\to 0} u^{s-1}\frac{b(u)u}{e^u - 1} = 0$$

where $u = z2^{-k}$. Assume first $s > 1$. Then

$$\lim_{u\to 0} u^{s-1}\frac{b(u)u}{e^u - 1} = 0$$

for any values of b_k, $k = 0, 1, \cdots$, assuming $b(0) < \infty$ (in our case $b(0) = b_0 = 0$). Let now $s \leqq 1$ and apply l'Hospital's rule $1 - s$ times. Then

$$0 = \lim_{u\to 0} u^{s-1}\frac{b(u)u}{e^u - 1} = \lim_{u\to 0}\frac{b(u)}{u^{1-s}} = \lim_{u\to 0}\frac{b^{(1)}(u)}{(1-s)u^{-s}} = \cdots = \lim_{u\to 0}\frac{b^{(1-s)}(u)}{(1-s)!};$$

hence (3.10) holds.

Sufficiency. Assume now that (3.10) is satisfied. By D'Alembert's criterion [11] the series is convergent if $\lim_{k\to\infty} \alpha_{k+1}/\alpha_k < 1$. Note that

$$\lim_{k\to\infty}\frac{\alpha_{k+1}}{\alpha_k} = \lim_{k\to\infty} 2^{-1}\frac{e^{z2^{-k}} - 1}{e^{z2^{-k-1}} - 1}2^{1-s}\frac{b(z2^{-k-1})}{b(z2^{-k})}. \tag{3.11}$$

But, by l'Hospital's rule

$$\lim_{k\to\infty} 2^{-1}\frac{e^{z2^{-k}} - 1}{e^{z2^{-k-1}} - 1} = \lim_{u\to 0} 2^{-1}\frac{e^u - 1}{e^{u/2} - 1} = 1; \tag{3.12}$$

hence by (3.11) and (3.12)

$$\lim_{k\to\infty}\frac{\alpha_{k+1}}{\alpha_k} = 2^{1-s}\lim_{u\to 0}\frac{b(u2^{-1})}{b(u)} = 2^{1-s}\lim_{u\to 0} 2^{-1}\frac{b'(u2^{-1})}{b'(u)}$$

$$= \cdots = 2^{-1}\lim_{u\to 0}\frac{b^{(2-s)}(u2^{-1})}{b^{(2-s)}(u)} \leqq 2^{-1} < 1;$$

hence the series is convergent. □

In order to find an explicit formula for $l(z)$, let us introduce the Bernoulli inverse relation. For a given sequence A_n, $n = 0, 1$, we define a sequence $\hat{A}_n$, $n = 0, 1, \cdots$, as in [17]

$$\hat{A}_n = \sum_{k=0}^{n}\binom{n}{k} B_k A_{n-k} \tag{3.13}$$

where B_k, $k = 0, 1, \cdots$, are Bernoulli numbers defined by [1]

$$\frac{z}{e^z - 1} = \sum_{k=0}^{\infty} B_k \frac{z^k}{k!}, \qquad |z| < 2\pi. \tag{3.14}$$

(In Appendix A we list some properties of Bernoulli numbers and Bernoulli polynomials which will be used in the further part of this paper.) Note also that by (3.13) and (3.14) the exponential generating function $\hat{A}(z)$ of $\hat{A}_n$ is given by [1], [17]

$$\hat{A}(z) = A(z)\frac{z}{e^z - 1} \tag{3.15}$$

and

$$A_n = \sum_{k=0}^{n}\binom{n}{k}(k+1)^{-1}\hat{A}_{n-k}$$

(by the above equation and (3.13) A_n and $\hat{A}_n$ are a pair of inverse relations). Then, we prove our first main result of this section.

THEOREM 3.4. *If* (3.10) *and the hypothesis of Corollary* 3.3 *hold, then the exponential generating function of* l_n *is given by*

$$l(z) = l_0 e^z + z^{(1-s)^+ - 1} l_s^*(e^z - 1) + b(z) + (e^z - 1)\sum_{k=1}^{\infty} 2^{-sk}\frac{b(z2^{-k})}{e^{z2^{-k}} - 1} \tag{3.16}$$

and

$$(3.17)\qquad l(z)=l_0e^z+z^{(1-s)^+-1}l_s^*(e^z-1)+b(z)+\frac{(e^z-1)}{z}\sum_{k=(2-s)^-}^{\infty}\frac{\hat{b}_kz^k}{k!(2^{k+s-1}-1)}$$

where

$$(3.18)\qquad l_s^*=\hat{l}_{(1-s)^+}-l_0\delta_{1,(1-s)^+}-l_0B_{(1-s)^+}$$

and $a^-=\min\{a,0\}$, δ_{nk} *is the Kronecker delta, while* $b(z)$ *is defined in* (3.5a).

Proof. Equation (3.16) follows directly from (3.1), (3.6), (3.8) and (3.9). We must only derive (3.18). But, by (3.6) and (3.15) $H(z)=\hat{l}(z)-l_0z-l_0z/(e^z-1)$. Note now that $H^*(z)$ given by (3.9) is a coefficient of the Taylor expansion of $H(z)$ at $z^{(1-s)^+}$.

To prove (3.17), consider the series in (3.16), Corollary 3.3 and (3.15). Note also that by (3.13) condition (3.10) is equivalent to $\hat{b}_0=\hat{b}_1=\cdots=\hat{b}_{1-s}=0$. Then the series in (3.16) is equal to

$$\sum_{k=1}^{\infty}2^{(1-s)k}\frac{b(z2^{-k})z2^{-k}}{e^{z2^{-k}}-1}=\sum_{k=0}^{\infty}2^{(1-s)k}\hat{b}(z2^{-k})=\sum_{k=1}^{\infty}2^{(1-s)k}\sum_{i=(2-s)^-}^{\infty}\frac{\hat{b}_i}{i!}z^i2^{-ki}$$

$$=\sum_{k=(2-s)^-}^{\infty}\frac{\hat{b}_kz^k}{k!}\sum_{i=1}^{\infty}2^{-i(k+s-1)}=\sum_{k=(2-s)^-}^{\infty}\frac{\hat{b}_kz^k}{k!(2^{k+s-1}-1)}$$

and by (3.10) the geometric series in the above formula is convergent. □

The second main result of this section is as follows:

THEOREM 3.5. *If* (3.10) *holds and the hypothesis of Corollary* 3.2 *is satisfied, then recurrence* (2.1) *possesses the following solution*:

$$(3.19)\qquad l_n=l_0+(1-\delta_{n0})l_s^*\frac{n!}{(n+s)!}+b_n+\frac{1}{n+1}\sum_{k=(2-s)^-}^{n}\binom{n+1}{k}\frac{\hat{b}_k}{2^{k+s-1}-1}$$

where

$$(3.20)\qquad b_0=0,\quad b_n=2^{-s}a_n-l_0(1-2^{-s})+g_n\chi_{(n\leqq N)},\quad n>0,$$

$$(3.21)\qquad \hat{b}_k=2^{-s}(\hat{a}_k-a_0B_k)-l_0(1-2^{-s})\delta_{k1}+(1-\delta_{k0})\sum_{i=1}^{\min\{k,N\}}\binom{k}{i}g_iB_{k-i}.$$

g_i *is given by* (3.5b) *and* χ_A *is a function equal to one if condition* A *is satisfied, and otherwise it is zero.*

Proof. Equation (3.19) follows directly from (3.17) by applying the multiplication formula for generating functions. □

Remarks. (i) Assumptions (a) and (b) from § 2 are relevant while (c) is not relevant for the above derivations. If a constant γ is any number, then (3.4) becomes

$$L(z)=\gamma L(z/2)(e^{z/2}+1)+b(z)$$

and using (3.6) together with $e^z-1=(e^{z/2}-1)(e^{z/2}+1)$ we obtain

$$H(z)=2\gamma H(z/2)+\frac{b(z)z}{e^z-1}$$

instead of (3.7). (Note that $b(z)$ here is defined slightly different than in (3.5a).) This form of the above functional equation is relevant to get a closed form solution for $H(z)$.

(ii) For (3.19) (more precisely: (3.21)) we must compute $\hat{a}_n$ for a given a_n, $n=0,1,\cdots$. For example, if $a_n=\binom{n}{r}q^n$, q is a constant and r is an integer, then using (A4) from Appendix A and (3.13) we find

$$\hat{a}_n = \sum_{k=0}^{n} \binom{n}{k} B_k \binom{n-k}{r} q^{n-k}$$

$$= \binom{n}{r} q^r \sum_{k=0}^{n-r} \binom{n-r}{k} B_k q^{n-r-k} = \binom{n}{r} q^r B_{n-r}(q)$$

where $B_{n-r}(q)$ is a Bernoulli polynomial (see (A1)). Hence

$$a_n = \binom{n}{r} q^n, \qquad \hat{a}_n = \binom{n}{r} q^r B_{n-r}(q). \tag{3.22}$$

For more Bernoulli inverse relations, see Riordan [17].

Example 1b. In that case we must substitute in (2.1) $c_0 = c_1 = 0$, $N = 1$, $s = 0$, $a_n = 1 - 2 \cdot 2^{-n}$. Then, by (3.5b), (3.18) and (3.20) $g_1 = 0$; $l_0^* = 0$; $b_k = 1 - 2^{1-k} + \delta_{k0}$. But $B_n(\frac{1}{2}) = -(1 - 2^{1-n})B_n$, (see (A9) in Appendix A) so $\hat{a}_n = 3B_n + \delta_{n1} - 2^{2-n}B_n$. Then, by (3.22) and the above $\hat{b}_n = 4B_n(1 - 2^n) + \delta_{n1}$ and by (3.19)

$$c_n = 1 - 2^{1-n} + \delta_{n0} + \frac{4}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k(1-2^{-k})}{2^{k-1}-1}.$$

Using now (A4) and (A8) after some algebra we obtain

$$c_n = 2 - \frac{4}{n+1} + 2\delta_{n0} + \frac{2}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{2^{k-1}-1}. \tag{3.23}$$

Example 2b. Consider first (2.5). Then, $a_n = 1$, $s = 0$, $N = 1$, and $q_1 = -1$, $l_0^* = 0$, $b_n = 1 - \delta_{n0} - \delta_{n1}$. Naturally, $\hat{a}_n = B_n + \delta_{n1}$ and $\hat{b}_n = -nB_{n-1}$ for $n \geqq 2$. Therefore, by (3.19) we find

$$t_n = 2 - \delta_{n0} - \delta_{n1} - \frac{1}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \binom{k}{1} \frac{B_{k-1}}{2^{k-1}-1}. \tag{3.24}$$

On the other hand, for (2.6) we have $a_n = 2^{-n}$, $s = 1$, $N = 1$, and then we compute $g_1 = 0.25$, $l_1^* = 0$, $b_n = 2^{-n-1} - 0.5 + 0.25\delta_{n1}$. Moreover, by (3.22) $\hat{a}_n = B_n(2^{1-n} - 1)$ and $\hat{b}_n = B_n(2^{-n} - 1) - 0.5\delta_{n1} + 0.25nB_{n-1}$. Hence, after some algebra

$$w_n = \frac{1}{n+1} + \frac{1}{4}\delta_{n1} + \frac{0.25}{n+1} \sum_{k=1}^{n} \binom{n+1}{k} \binom{k}{1} \frac{B_{k-1}}{2^k-1}. \tag{3.25}$$

Example 3b. For (2.8) we assume $s = 0$, $N = 1$, $a_n = 2 + n2^{-n} - 2^{-n}$. Then $g_1 = -2$, $l_0^* = 1$ and $b_n = a_n - \delta_{n0} - 2\delta_{n1}$. Using (3.22) we show that $\hat{a}_k = 2(B_n + \delta_{n1}) + 0.5kB_{k-1}(1/2) - B_k(1/2)$ and by (A9) we obtain $\hat{b}_k = B_k - 1.5kB_{k-1} + 2kB_{k-1}(2^{-k} - 2^{-1}) - B_k(2^{1-k} - 1) + 1.5\delta_{n1}$. Hence, by the above and (A8) we find

$$\begin{aligned} t_n = {} & 4.5 - 2.5\delta_{n0} - 2.5\delta_{n1} - \frac{2}{n+1} + \frac{1}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{2^{k-1}-1} \\ & - \frac{1.5}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \binom{k}{1} \frac{B_{k-1}}{2^{k-1}-1}. \end{aligned} \tag{3.26}$$

Recurrence (2.9) is equivalent to (2.6) so the solution is given by (3.25).

4. Approximations. In this section we present an approximation for the exponential generating function $l(z)$ for small values of z and an asymptotic approximation of l_n for large values of n.

Let us start with the small value approximation of $l(z)$. Such an approximation might be very useful in practice if one is more interested in $l(z)$ than l_n. For example, in determining maximum throughput for a conflict resolution algorithm we must optimize a ratio given by (2.7), where exponential generating functions are involved. It turns out that the optimal value of z is rather small, and hence the discussed approximation is applied. Assume now $z < \beta$, β is a small real value and consider (3.17). Then, we find

$$(4.1)\quad l(z) = l_0 e^z + z^{(1-s)^+ - 1} l_s^*(e^z - 1) + b(z) + (e^z - 1) \sum_{k=(2-s)^-}^{M} \frac{\hat{b}_k z^{k-1}}{k!(2^{k+s-1} - 1)} + O(z^{M+1})$$

where $b(z)$, l_s^*, $\hat{b}_k$ are given by (3.5a), (3.18) and (3.21), respectively, while $M > (2 - s)^-$. The value of M determines the quality of the approximation.

Hereafter we deal only with the asymptotic analysis of l_n computed as in (3.19). Naturally, the problem is to find an approximation for the sum in (3.19), and further we restrict our considerations to that sum. Let

$$(4.2)\qquad S(n, s, b_n) = \frac{1}{n+1} \sum_{k=t}^{n} \binom{n+1}{k} \frac{\hat{b}_k}{2^{k+s-1} - 1}$$

where t is an integer. In our case $t = (2 - s)^-$. According to (3.21) $\hat{b}_k$ consists of three terms; however, for asymptotic analysis of (4.2) the first one is the most difficult to handle, since it includes $\hat{a}_k$. Furthermore, we restrict our considerations to a wide class of a_k such that the other terms of (3.21) will be automatically included in the analysis. Therefore, let a_n be given by (3.22), that is

$$(4.3)\qquad a_n = \binom{n}{r} q^n, \qquad \hat{a}_n = \binom{n}{r} q^r B_{n-r}(q)$$

where r is an integer while $q > 0$. For $q = 1$ we obtain, as a special case, the other terms of (3.21).

Under the above assumption we deal with the asymptotic approximation of the following

$$(4.4)\qquad S(n, r, s) = \frac{q^r}{n+1} \sum_{k=t}^{n} \binom{n+1}{k} \binom{k}{r} \frac{B_{k-r}(q)}{d^{k+s-1} - 1}$$

where $d > 1$, $t > 1 - s$. In our case $d = 2$. Note that applying the geometric series formula to the denominator of (4.4) one finds

$$(4.5)\qquad S(n, r, s) = \binom{n+1}{r} \frac{q^r}{n+1} \sum_{j=1}^{\infty} d^{-j(r+s)} \sum_{k=m}^{n-r} \binom{n+1-r}{k} B_k(q) d^{-j(k-1)}$$

where $m = \max\,\{t, r\} - r$. In our case $t = (2 - s)^-$, and

$$(4.6)\qquad m = \begin{cases} 2 - s - r, & s \leqq \max\,\{1, 2 - r\}, \\ 0 & \text{otherwise.} \end{cases}$$

Let us now consider the inner sum in (4.5) divided by n. Then

$$(4.7)\quad \frac{1}{n} \sum_{k=m}^{n-r} \binom{n+1-r}{k} B_k(q) d^{-j(k-1)} = \sum_{k=m}^{n-r} \binom{n+1-r}{k} \frac{1}{n^k} B_k(q) \left(\frac{n}{d^j}\right)^{k-1}.$$

But [13]

$$(4.8)\qquad \binom{n+1-r}{k} \frac{1}{n^k} = \frac{1}{k!}[1 + O(n^{-1})]$$

and let $x = nd^{-j}$. Equation (4.8) suggests the following approximation of (4.7) for large values of n

$$\sum_{k=m}^{n-r} \binom{n+1-r}{k} \frac{1}{n^k} B_k(q) x^{k-1} \approx \sum_{k=m}^{\infty} \frac{B_k(q) x^{k-1}}{k!}.$$

From (A1) we know that for $|x| < 2\pi$

$$\sum_{k=m}^{\infty} \frac{B_k(q)}{k!} x^{k-1} = \frac{e^{qx}}{e^x - 1} - \sum_{k=0}^{m-1} \frac{B_k(q) x^{k-1}}{k!}. \tag{4.9}$$

Therefore, we approximate $S(n, r, s)$ by $T(n, r, s)$ where

$$T(n,r,s) = \binom{n+1}{r} \frac{nq^r}{n+1} \sum_{j=1}^{\infty} d^{-j(r+s)} \left[\frac{e^{qx}}{e^x - 1} - \sum_{k=0}^{m-1} \frac{B_k(q) x^{k-1}}{k!} \right]. \tag{4.10}$$

We prove that

THEOREM 4.1. *For any values of r and s*

$$\delta(n,s) \overset{\text{def}}{=} T(n,r,s) - S(n,r,s) = O(n^{-s-1}). \tag{4.11}$$

Proof. Let $\delta(n, s) = \delta_1(n, s) + \delta_2(n, s)$ where $\delta_1(n, s)$ is computed for $x < 1$ ($nd^{-j} < 1$) while $\delta_2(n, s)$ for $x > 1$. We first evaluate $\delta_1(n, s)$. Then for $d^j > n$ one finds

$$\begin{aligned} \delta_1(n,s) = O(n^r) \sum_{j=\log_d n}^{\infty} d^{-j(r+s)} & \left| \sum_{k=m}^{n-r} B_k(q) x^{k-1} \right. \\ & \left. \times \left[\frac{1}{k!} - \binom{n+1-r}{k} \frac{1}{n^k} \right] + \sum_{k=n-r+1}^{\infty} \frac{B_k(q) x^{k-1}}{k!} \right| \\ \leqq O(n^r) \sum_{j=\log_d n}^{\infty} & d^{-j(r+s)} O(n^{-1} x^{m-1}). \end{aligned} \tag{4.12}$$

The inequality in (4.12) comes from the fact that the second term in (4.12) represents a remainder of a convergent series and we can make it as small as we want for large values of n, so the first term in (4.12) is a leading factor. Therefore,

$$\begin{aligned} \delta_1(n,s) &\leqq \sum_{j=\log_d n}^{\infty} d^{-j(r+s)} O(n^{r-1} n^{m-1} d^{-j(m-1)}) \\ &= \sum_{j=\log_d n}^{\infty} d^{-j(r+s+m-1)} O(n^{r+m-2}) = O(n^{-s-1}) \end{aligned}$$

since under (4.6) the above geometric series is convergent.

Assume now $x > 1$, i.e., $d^j < n$. For simplicity we also assume that $m = 0$ and $q = 1$. Then the finite sum in (4.5) may be rewritten in the presence of (A8) as

$$A \overset{\text{def}}{=} \frac{1}{n+1-r} \sum_{k=0}^{n-r} \binom{n+1-r}{k} B_k d^{-j(k-1)} = \frac{1}{d^{j(n-r)}} \sum_{k=1}^{d^j - 1} k^{n-r}.$$

But for any $a > 1$

$$\sum_{k=1}^{a-1} (ka^{-1})^n \leqq a^{-n} \int_1^{a-1} x^n dx \leqq (n+1)^{-1} a^{-1} (1 - a^{-1})^{n+1};$$

hence $A \leqq O(x^{-1}(1 - d^{-j})^n)$. Using the following well-known inequalities [13]

$$\frac{1}{e^x - 1} > \frac{e^{-x}}{x}, \quad x > 0, \quad 1 - d^{-j} < e^{-d^{-j}},$$

we may evaluate $\delta_2(n, r)$ as follows:

$$\delta_2(n,r) \leqq O(n^r) \sum_{j=1}^{\log_d n} d^{-j(r+s)} \left[\frac{1}{x}(1-d^{-j})^n - \frac{1}{e^x - 1} \right]$$

$$\leqq O(n^r) \sum_{j=1}^{\log_d n} d^{-j(r+s)} O\left(\frac{e^{-x}}{x} \right) \leqq O(\exp\{-[nd + (r+s)\ln d - r \ln n - \ln \log_d n]\}),$$

which may be made as small as we need for large n, e.g., $O(n^{-s-1})$ as required. □

By Theorem 4.1 the problem of computing $S(n, r, s)$ is reduced to finding $T(n, r, s)$ given by (4.10). We apply the Mellin transform method [9], [12], [15]. In Appendix B we prove that for an odd integer c, real x, $q > 0$ and $\operatorname{Re} z > c/2$

$$(4.13) \qquad I(c,q) \overset{\text{def}}{=} \frac{1}{2\pi i} \int_{c/2 - i\infty}^{c/2 + i\infty} \zeta(z,q)\Gamma(z) x^{-z} dz = \frac{e^{x(1-q)}}{e^x - 1} - \sum_{k=0}^{(1-c)/2} \frac{B_k(1-q)}{k!} x^{k-1}$$

where $\zeta(z, q)$ is the generalized zeta function (see Appendix A) and $\Gamma(z)$ is the gamma function [1], [5]. Then (4.10) is equal to

$$(4.14) \qquad T(n,r,q) = \binom{n+1}{r} \frac{nq^r}{n+1} \sum_{j=1}^{\infty} d^{-j(r+s)} I(3-2m, 1-q+\delta_{q1})$$

and we restrict the range of q to $0 \leqq q \leqq 1$. The case $q = 1$ needs some additional considerations, therefore δ_{q1} appears in (4.14) as shown in Appendix B. Note that $x = nd^{-j}$, and then after some algebra one proves that for $\operatorname{Re} z < r + s$

$$(4.15) \qquad T(n,r,q) = [1 + O(n^{-1})] \frac{q^r}{r!} \int_{(3/2-m)} \frac{\zeta(z, 1-q+\delta_{q1})\Gamma(z) n^{r-z}}{d^{r+s-z} - 1} dz$$

where $\int_{(c)}$ stands for $1/2\pi i \int_{c-i\infty}^{c+i\infty}$.

The calculation of the contour integral in (4.15) is routine, and is equal to minus the sum of the residues of the function under integral to the right of the line of integration [9]. Three types of singularities must be taken into account:

(i) *Zeros of the denominator*, that is, $d^{r+s-z} - 1 = 0$. The roots of this equation are equal to $z_k = r + s + 2\pi i k/\ln d$, $k = 0, \pm 1, \cdots$,

(ii) *Singular point of the zeta function* at $z = 1$,

(iii) *Singular points of the gamma function* at $z = -m$, where m is a nonnegative integer.

The number of singularities we must consider for evaluation of (4.15) depends on the position of the line of integration, that is, it depends on the values of r and s. The most difficult to handle is a double pole, which might occur if a zero of the denominator coincides with singular point of the zeta function or with the gamma function. We prove

PROPOSITION 4.2. *Let* $M = r + s$.

(i) *If* $M = r + s < 1$ *then*

$$S(n,r,s) = -\frac{n^{-s}q^r}{r!} \left\{ \frac{(-1)^{-M+1}}{(-M)! \ln d} \left[\frac{B_{1-M}(1-q+\delta_{q1})}{(1-M)} (\ln n - \psi(1-M) - 0.5 \ln d) \right.\right.$$

$$(4.16) \qquad \left. + \zeta'(M, 1-q+\delta_{q1}) \right]$$

$$\left. + \sum_{\substack{l=-1 \\ l \neq -M}}^{m-2} \frac{(-1)^{l+1}}{(l+1)!} \frac{B_{l+1}(1-q+\delta_{q1})}{d^{M+l} - 1} n^{l+M} - f_M(n) \right\}$$

$$+ O(n^{-s-1});$$

(ii) *If $M = r + s = 1$, then*

$$S(n,r,s) = -\frac{n^{-s}q^r}{r!}\left\{\ln^{-1} d\,[\ln n + \gamma + \psi(1-q+\delta_{q1}) - 0.5\ln d] \right.$$

$$\left. + \sum_{l=0}^{m-2} \frac{(-1)^{l+1}}{(l+1)!}\frac{B_{l+1}(1-q+\delta_{q1})}{d^{M+l}-1} n^{M+l} - f_M(n)\right\} + O(n^{-s-1}); \tag{4.17}$$

(iii) *If $M > 1$ and $m > \frac{1}{2}$ then*

$$S(n,r,s) = -\frac{n^{-s}q^r}{r!}\left\{\sum_{l=-1}^{m-2} \frac{(-1)^{l+1}}{(l+1)!}\frac{B_{l+1}(1-q+\delta_{q1})}{d^{M+1}-1} n^{M+1} \right.$$

$$\left. - f_M(n) - \zeta(M, 1-q+\delta_{q1})(M-1)!\ln^{-1} d\right\} + O(n^{-s-1}); \tag{4.18}$$

(iv) *If $M > 1$ and $m < \frac{1}{2}$ then*

$$S(n,r,s) = \frac{n^{-s}q^r}{r!}\{\zeta(M, 1-q+\delta_{q1})(M-1)!\ln^{-1} d + f_M(n)\} + O(n^{-s-1}) \tag{4.19}$$

where

$$f_M(n) = \frac{1}{\ln d}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \zeta(M + 2\pi ik/\ln d)\Gamma(M + 2\pi ik/\ln d)\exp(-2\pi ik \log_d n) \tag{4.20}$$

and $\psi(x)$ is the psi function [1].

Proof. See Appendix C. □

The function $f_M(n)$ may be safely ignored for practical purposes, since numerical analysis shows that values of the function are very small in comparison with the leading component of (4.16)–(4.19). In fact, it is not difficult to prove that $f_M(n) = f_M(dn)$ (e.g. $f_M(n)$ is a periodic function of $\log_d n$) and f_M is bounded. The last follows from the following well-known formula [19]

(i) $$|\exp(iy)| \leqq 1, \qquad y - \text{real},$$

(ii)

$$\zeta(s+iy, q) = \begin{cases} O(1) & \text{for } s > 1, \\ O(y^{1/2-s}) & \text{for } s < 0, \end{cases}$$

(iii) for any nonnegative integer, s,

$$|\Gamma(s+iy)|^2 = \frac{\pi}{y\sinh y}\prod_{j=0}^{s-1}(j^2+y^2),$$

$$|\Gamma(-s+iy)|^2 = \frac{\pi^2}{y\sinh y}\left[\prod_{j=1}^{s}(j^2+y^2)\right]^{-1}$$

and $\sinh y = O(e^{|y|})$.

Example 1c. To evaluate (3.23) we put $s = 0$, $r = 0$, $q = 1$, $m = 2$ in (4.16) and after some algebra we obtain

$$c_n = \lg(n) - \frac{\ln\pi - \gamma}{\ln 2} + \frac{1}{2} + f_0(n) + O(n^{-1}), \qquad n \geqq 1 \tag{4.21}$$

where $f_0(n)$ is given by (4.20) with $M = 0$ (see also [12]).

Example 2c. For (3.24) we assume $s = 0$, $r = 1$, $m = 1$ and by (4.17) we find

$$t_n = 1.5 - \delta_{n1} + \lg n + f_1(n) + O(n^{-1}), \qquad n \geqq 1 \tag{4.22}$$

where $\lg n = \log_2 n$. For (3.25) we assume $s = 1$, $r = 1$, $m = 0$, hence by (4.19) and (A16) we obtain

$$w_n = \frac{1}{n+1} + \frac{1}{4}\delta_{n1} + \frac{\pi^2}{24 \ln 2} \cdot \frac{1}{n} + \frac{1}{4n} f_2(n) + O(n^{-2}). \tag{4.23}$$

However, in order to compute the maximum throughput of the algorithm (Eq. (2.7)) we need the exponential generating function of t_n and w_n for small values of z. But by (4.7) one finds

$$t(z) = \frac{z}{2} - \frac{z^2}{36} + \frac{z^4}{24 \cdot 450} + O(z^6), \tag{4.24a}$$

$$w(z) = 1 - \frac{z}{6} + \frac{z^2}{84} - \frac{z^4}{720 \cdot 31} + O(z^8) \tag{4.24b}$$

where z is a real number, which might be optimized to get maximum throughput. For details see [16].

Example 3c. By (3.26), the two sums might be evaluated either as in Example 1c or as in Example 2c. We immediately obtain

$$t_n = 3 + 0.5 \lg (n^4/\pi) + \frac{\gamma}{2 \ln 2} + O(n^{-1}) \tag{4.25}$$

while w_n is given by (4.25).

5. Conclusions. In this paper a *linear recurrence with full history* was considered. We have found a closed-form solution of the recurrence, and in addition the generating function of the solution was computed. We have also established two approximations: a small value approximation for the generating function and an asymptotic approximation for l_n. The analysis was illustrated by three examples of some importance in practice. In future research assumption (a) from § 2 should be relaxed. Note also that for an exact solution of (2.3) and (2.4) we should consider (2.1) with nl_{n-1} added to the LHS of (2.1). Finally, our basic recurrence (2.1) with a_n instead of $2^n a_n$ should be also analyzed.

Appendix A. Bernoulli polynomials and the Riemann zeta function. We list below some properties of Bernoulli polynomials and the Riemann zeta function which are used often in this paper. Details may be found in [1], [3], [5], [9], [11], [14].

Bernoulli numbers B_n and Bernoulli polynomials $B_n(x)$.

DEFINITION.

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} B_n(x) \frac{t^n}{n!}, \qquad |t| < 2\pi, \tag{A1}$$

$$B_n = B_n(0). \tag{A2}$$

Properties.

$$B_n(x+1) = B_n(x) + nx^{n-1}, \tag{A3}$$

$$B_n(x+h) = \sum_{k=0}^{n} \binom{n}{k} B_k(x) h^{n-1}, \tag{A4}$$

$$B_n(1-x)=(-1)^n B_n(x), \tag{A5}$$

$$B_n=\sum_{k=0}^{n}\binom{n}{k}B_k-\delta_{n1}, \tag{A6}$$

$$\sum_{k=1}^{m}k^n=\frac{B_n(m+1)-B_{n+1}}{n+1}, \qquad n,m=1,2,\cdots, \tag{A7}$$

$$\int_a^x B_n(t)dt=\frac{B_{n+1}(x)-B_{n+1}(a)}{n+1}, \tag{A8}$$

$$B_n\left(\frac{1}{2}\right)=B_n(2^{1-n}-1). \tag{A9}$$

Generalized zeta function.

$$\zeta(z,q)=\sum_{n=0}^{\infty}\frac{1}{(q+n)^z}, \quad \operatorname{Re} z>1, \quad q\neq 0,-1,-2,\cdots. \tag{A10}$$

Riemann zeta function.

$$\zeta(z)=\zeta(z,1). \tag{A11}$$

Properties.

$$\zeta(0,q)=\frac{1}{2}-q, \tag{A12}$$

$$\lim_{z\to 1}\left[\zeta(z,q)-\frac{1}{s-1}\right]=-\psi(q) \tag{A13}$$

where $\psi(z)=\Gamma'(z)/\Gamma(z)$.

$$\left.\frac{d\zeta(z,q)}{dz}\right|_{z=0}=\ln\Gamma(q)-\frac{1}{2}\ln 2\pi, \tag{A14}$$

$$\zeta(-n,q)=-\frac{B_{n+1}(q)}{n+1}, \qquad n=0,2,3,\cdots, \tag{A15}$$

$$\zeta(2m)=\frac{(2\pi)^{2m}}{2(2m)!}|B_{2m}|, \qquad m=1,2,\cdots. \tag{A16}$$

Appendix B: Mellin transform. Let us compute the following integral:

$$I(c,q)=\frac{1}{2\pi i}\int_{c/2-\infty}^{c/2+i\infty}\zeta(z,q)\Gamma(z)x^{-z}dz, \qquad \operatorname{Re} z>\frac{c}{2} \tag{B1}$$

where $q>0$ and c is an odd integer, while x is real. To evaluate the integral, we use the residue method. A path of integration goes from $(c/2+iN)$ to $(c/2+iN-M)$ to $(c/2-iN-M)$ to $(c/2-iN)$ to $(c/2+iN)$. Using properties of zeta and gamma functions ([9], [19]) one easily proves that the integral over horizontal lines and left vertical line vanishes where $N, M\to\infty$. Therefore, $I(c,q)$ is equal to the sum of residues left of the line $(c/2-i\infty, c/2+i\infty)$.

Case A: $c \leqq 0$. Then the only singularities of the integrand are poles of the gamma function, that is, nonpositive integers smaller than $c/2$. Hence, noting that for $z = -k$ ($k \geqq 0$) the residue of the gamma function is equal to $(-1)^k/k!$ [9], we obtain

$$I(c,q) = \sum_{k=(1-c)/2}^{\infty} \zeta(-k,a)\frac{(-1)^k x^k}{k!}.$$

But by (A15), (A5) and (A1) for $|x| < 2\pi$ we find

$$\text{(B2)} \qquad I(c,q) = \frac{e^{x(1-q)}}{e^x - 1} - \sum_{k=0}^{(1-c)/2} \frac{B_k(1-q)}{k!} x^{k-1}, \qquad |x| < 2\pi$$

where $B_k(x)$ is Bernoulli polynomial.

Case B: $c > 0$. In that case all nonpositive integers are singularities of the gamma function, and in addition for $c = 1$ there is a simple singularity at $z = 1$ of zeta function. Therefore,

$$I(c,q) = \sum_{k=0}^{\infty} \zeta(-k,q)\frac{(-1)^k x^k}{k!} + (1 - \delta_{c1})x^{-1}$$

since the residue at $z = 1$ of the zeta function is equal to one. As above using (A15), (A5) and (A1) we finally obtain

$$\text{(B3)} \qquad I(c,q) = \frac{e^{x(1-q)}}{e^x - 1} - x^{-1}\delta_{c,1}, \qquad |x| < 2\pi.$$

By analytical continuation we prove that (B2) and (B3) hold for all real x. Hence, for any odd integer c we find

$$\text{(B4)} \qquad I(c,q) = \frac{e^{x(1-q)}}{e^x - 1} - \sum_{k=0}^{(1-c)/2} \frac{B_k(1-q)}{k!} x^{k-1}$$

where the sum in (B4) is assumed to be zero if the upper index is smaller than the lower index in the sum symbol. Moreover, for $q = 0$ it is easy to show (using the fact: $B_n(0) = B_n(1)$ for $n > 1$ and $B_1(1) = -B_1(0)$) that

$$\text{(B5)} \qquad I(c,1) = \frac{1}{e^x - 1} - \sum_{k=0}^{(1-c)/2} \frac{B_k}{k!} x^{k-1}$$

and then $\zeta(z, q)$ in (B1) becomes the Riemann zeta function $\zeta(z) = \zeta(z, 1)$.

Appendix C: Proof of Proposition 4.2. Let us evaluate the following integral:

$$\text{(C1)} \qquad J(n,N,M) = \frac{1}{2\pi i}\int_{c/2 - i\infty}^{c/2 + i\infty} \frac{\zeta(z,q)\Gamma(z)n^{N-z}}{d^{M-z} - 1}\,dz$$

where c is an odd integer, $q > 0$, N, M-integers, and Re $z < M$, $c/2 < M$.

In the evaluation of the integral we use the same method as in Appendix B; however, this time the path of integration is right to the line $(c/2 - i\infty,\ c/2 + i\infty)$. By the same arguments as above we can show that the integral is minus the sum of residues on the right of the line of integration.

Let $g(z)$ be a function under integral. We must consider four cases depending on the value of M and c (note that $c/2 < M$).

Case A: $M < 1$. In that case singularities of $g(z)$ are as follows:

(a) *For the gamma functions,* all nonpositive integers in the interval $[(c + 1)/2, 0]$, that is, $z = -m$, $m = 0, 1, \cdots, -(c+1)/2$;

(b) *For the zeta function* at $z = 1$;

(c) *Zeros of the denominator* in (C1), that is, $z_k = M + 2\pi ik/\ln d$, $k = 0, \pm 1, \pm 2, \cdots$.

Then the residues of $g(z)$ are equal to the following:

(i) For $z = -m$, $m \neq M$, $m = 0, 1, \cdots, -(c + 1)/2$ by (C1) and (A15)

$$\text{(C2)} \qquad \operatorname*{res}_{z=-m\neq M} g(z) = \zeta(-m, q)\frac{(-1)^m}{m!}\frac{n^{N+m}}{d^{M+m}-1} = \frac{(-1)^{m+1}}{(m+1)!}\frac{B_{m+1}(q)n^{N+m}}{d^{M+m}-1};$$

(ii) For $z = 1$ ([9], [19])

$$\text{(C3)} \qquad \operatorname*{res}_{z=1} g(z) = \frac{n^{N-1}}{d^M - 1};$$

(iii) For $z_k = M + 2\pi ik/\ln d$, $k = \pm 1, \pm 2, \cdots$, $(z_k \neq M)$

$$\text{(C4)} \qquad \begin{aligned} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \operatorname{res} g(z_k) &= -n^{N-M}\frac{1}{\ln d}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \zeta(M + 2\pi ik/\ln d) \\ &\quad \times \Gamma(M + 2\pi ik/\ln d)\exp[-2\pi ik \log_d n] \\ &= -n^{N-M} f_M(n) \end{aligned}$$

where $f_M(n)$ is defined as in (4.20).

(iv) For $z = M$

This is the most difficult to handle, since $z = M$ is double pole of $g(z)$ (gamma function and the denominator of (C1)). To find the residue, we use the following expansions of the functions under the integral at $z = M$ (let $w = z + M$) ([5], [9]):

$$\text{(C5a)} \qquad \zeta(z, q) = -\frac{B_{1-M}(q)}{1-M} + w\zeta'(M, q) + O(w^2),$$

$$\text{(C5b)} \qquad \Gamma(z) = w^{-1}\frac{(-1)^{-M}}{(-M)!} + \frac{(-1)^{-M}}{(-M)!}\psi(1-M) + O(w),$$

$$\text{(C5c)} \qquad n^{N-z} = n^{N-M} - w\, n^{N-M}\ln n + O(w^2),$$

$$\text{(C5d)} \qquad \frac{1}{d^{M-z}-1} = -w^{-1}\frac{1}{\ln d} - \frac{1}{2} + O(w).$$

The residue at $z = M$ is the coefficient of w^{-1} in the product of (C5a)–(C5d). After some algebra we find that

$$\text{(C6)} \quad \operatorname*{res}_{z=M} q(z) = n^{N-M}\frac{(-1)^{1-M}}{(1-M)!}\left\{B_{1-M}(q)\left(\log_d n - \frac{1}{2} - \psi(1-M)/\ln d\right) + \zeta'(M, q)/\ln d\right\}$$

where $\psi(x)$ is the psi function [1], [5] and $\zeta'(x, q)$ denotes the derivative of the zeta function for $z = x$. For example ([1], [5])

$$\zeta'(0, q) = \ln\Gamma(q) - \tfrac{1}{2}\ln 2\pi.$$

For other values of $\zeta'(x, q)$ see [3].

Finally, taking into account (C2)–(C6) we find that for $c/2 < M < 1$

$$\begin{aligned} J(n,N,M)n^{M-N} = &-\frac{(-1)^{1-M}}{(1-M)!} \\ &\times\left\{B_{1-M}(q)\left[\log_d n - \frac{1}{2} - \psi(1-M)/\ln d\right] + \zeta'(M,q)/\ln d\right\} \\ &-\sum_{\substack{m=-1 \\ m\neq -M}}^{-(c+1)/2} \frac{(-1)^{m+1}}{(m+1)!}\frac{B_{m+1}(q)n^{m+M}}{d^{M+m}-1} + f_M(n) \end{aligned} \tag{C7}$$

where $f_M(n)$ is defined in (C4).

Case B: $M = 1$. In that case we have the same singularities as above; however, now $z = 1$ is a double pole of the denominator and the zeta function. Hence, (C2) holds for all nonnegative integers $m \in [0, -(c + 1)/2]$, and (C4) holds for $M = 1$. The only problem is the double pole at $z = 1$. But denoting $w = z - 1$ and using expansions (C5c), (C5d) together with [5], [9]

$$\zeta(z,q) = w^{-1} - \psi(q) + O(w), \qquad \Gamma(z) = 1 - \gamma w + O(w^2),$$

we find that

$$\operatorname*{res}_{z=1} q(z) = n^{N-1}\left(\log_d n + \frac{\gamma + \psi(q)}{\ln d} - \frac{1}{2}\right).$$

Therefore,

$$\begin{aligned} J(n,N,M)n^{1-M} = &-\left(\log_d n + \frac{\gamma+\psi(q)}{\ln d} - \frac{1}{2}\right) \\ &-\sum_{m=0}^{-(c+1)/2} \frac{(-1)^{m+1}}{(m+1)!}\frac{B_{m+1}(q)n^{n+1}}{d^{M+m}-1} + f_1(n) \end{aligned} \tag{C8}$$

where $f_1(n)$ is defined in (C4) for $M = 1$.

Case C: $M > 1$ *and* $c/2 < 1$. We have now the same singularities as before, but there is no double pole. Therefore, by the same arguments as above we find

$$J(n,N,M)n^{M-N} = -\sum_{m=-1}^{-(c+1)/2} \frac{(-1)^{m+1}}{(m+1)!}\frac{B_{m+1}(q)n^{m+M}}{d^{M+m}-1} + \frac{\zeta(M,q)(M-1)!}{\ln d} + f_M(n). \tag{C9}$$

Case D: $M > 1$ *and* $c/2 > 1$. In that case only zeros of the denominator are poles of the function under the integral. Noting that for $z_0 = M$

$$\operatorname*{res}_{z_0=M} q(z) = \frac{\zeta(M,q)(M-1)!n^{N-M}}{\ln d},$$

we find that

$$J(n,N,M)n^{M-N} = \zeta(M,q)(M-1)!/\ln d + f_M(n) \tag{C10}$$

where $f_M(n)$ is defined in (C4).

In order to prove (4.16)–(4.19) we must assume in (C7), (C8), (C9) and (C10) $N = r$ and $M = r + s$, and take into account (4.15).

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] T. BERGER, *Poisson multiple-access contention with binary feedback*, IEEE Trans. Inform. Theory, IT-30 (1985), pp. 745–751.

[3] B. BERNDT, *Ramanujan's Notebooks, Part* I, Springer–Verlag, Berlin, New York, 1985.

[4] J. CAPETANAKIS, *Tree algorithms for packet broadcast channels*, IEEE Trans. Inform. Theory, IT-25 (1979), pp. 505–515.

[5] A. ERDELYI, *Higher Transcendental Functions, Vol.* I, McGraw-Hill, New York, 1953.

[6] PH. FLAJOLET AND N. SAHEB, *The complexity of generating an exponentially distributed variate*, J. Algorithms, 7 (1986), pp. 463–488.

[7] R. GALLAGER, *Conflict resolution in random access broadcast networks*, Proc. AFOSR Workshop in Communication Theory and Applications, 1978, pp. 74–76.

[8] G. GONNET, *Handbook of Algorithms and Data Structures*, Addison–Wesley, Reading, MA, 1984.

[9] P. HENRICI, *Applied and Computational Complex Analysis, Vol.* 2, John Wiley, New York, 1977.

[10] M. HOFRI, *Stack algorithms for collision-detecting channels and their analysis: A limited survey*, Proceedings of Intern. Seminar on Modeling and Performance Evaluation Methodology, Vol. 1, Paris, 1983, pp. 53–78.

[11] K. KNOPP, *Theory and Application of Infinite Series*, Hafner, New York, 1949.

[12] D. KNUTH, *The Art of Computer Programming, Vol.* 3. *Sorting and Searching*, Addison–Wesley, Reading, MA, 1973.

[13] D. MITRINOVIC, *Analytical Inequalities*, Springer-Verlag, Berlin-New York, 1970.

[14] N. NIELSEN, *Nombres de Bernoulli*, Paris, 1923.

[15] W. SZPANKOWSKI, *Analysis of a recurrence equation arising in stack-type algorithms for collision-detecting channels*, Proc. Internat. Seminar on Computer Networking and Performance Evaluation, Tokyo, 1985, pp. 399–412.

[16] ———, *An analysis of a contention resolution algorithm—Another approach*, Acta Informatica, to appear.

[17] J. RIORDAN, *Combinatorial Identities*, John Wiley, New York, 1968.

[18] B. TSYBAKOV AND V. MIKHAILOV, *Random multiple packet access: Part-and-Try algorithm*, Problems of Information Transmission, 16 (1980), pp. 305–317.

[19] E. WHITTAKER AND G. WATSON, *A Course of Modern Analysis*, Cambridge Univ. Press, Cambridge, 1935.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

011

ALGEBRAIC METHODS APPLIED TO NETWORK RELIABILITY PROBLEMS*

DOUGLAS R. SHIER† AND DAVID E. WHITED†

Abstract. An algebraic structure underlying network reliability problems is presented for determining the 2-terminal reliability of directed networks. An iterative algorithm is derived from this algebraic perspective to solve the (s, j)-terminal reliability problem simultaneously for all nodes j. In addition to providing an exact answer (in the form of a reliability polynomial), the algorithm also yields a nondecreasing sequence of approximate solutions guaranteed to be lower bounds on the exact solution. Empirical results, presented for two different implementations of the algorithm, show that useful approximate solutions can be obtained in a reasonable amount of computation time.

Key words. algorithm, directed graphs, reliability

AMS(MOS) subject classifications. 05C20, 62N05, 94C15

1. Introduction. The problem of determining the reliability of an existing or proposed communication system has received considerable attention in the engineering, statistical, and operations research literature [1], [5], [9]. For example, it is important to assess the probability that a message sent from a given source arrives at its destination, when the components comprising the system are subject to failure. Unfortunately, most reliability problems of any substance are now known to be NP-hard or #P-complete [4], [17], [18], [27]. As a result, researchers have focused on special network structures (where polynomial-time algorithms are possible), or have resorted to simulation.

A number of special classes of *undirected* networks have recently been analyzed with success. Polynomial-time algorithms are now available for calculating certain reliability measures in series-parallel [21], inner-cycle-free [14], inner-four-cycle-free [14], and cube-free [15] planar graphs. Provan [16] has shown, however, that the problem of determining source-to-terminal reliability remains #P-complete for the general class of planar graphs. In order to analyze more complex network topologies, the idea of pivotal decomposition [5] together with polygon-to-chain reductions [28] can be used to decompose the original problem into smaller subproblems.

Similar results and tools are not as available in the case of *directed* networks. The only significant types of directed networks that are known to admit a polynomial-time algorithm are "basically-series-parallel" networks [2], [3]. Also, unlike the case for undirected networks, certain simplifications are not available when carrying out pivotal decomposition in the directed case [1]. Nor does there exist an "optimal" factoring algorithm, such as that demonstrated for undirected networks [19].

This paper exploits the underlying algebraic structure of network reliability problems to produce a general iterative algorithm, applicable to arbitrary directed networks. While not polynomially-bounded, it is able to generate reasonable approximations to exact network reliability with a modest amount of computation.

2. Algebraic structure. Suppose that $G = (N, E)$ is a directed network with node set N and edge set E, in which nodes do not fail but edges fail independently of one another. The *reliability* of edge e (the probability that edge e functions) is denoted by p_e.

* Received by the editors August 14, 1985; accepted for publication (in revised form) August 21, 1986. This research was supported by the U.S. Air Force Office of Scientific Research under grant AFOSR-85-0154.

† Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29634.

Nodes s and t designate the specified source and terminal of G, and we are interested in calculating the 2-terminal reliability $R_{st}(G)$:

$$R_{st}(G) = \Pr\{\text{there exists a functioning path from } s \text{ to } t \text{ in } G\}.$$

Associate with each edge $k \in E$ a *variable* x_k. Then the *reliability polynomial* $F_{st}(\mathbf{x}) = F_{st}(x_1, \cdots, x_r)$ associated with s and t is a polynomial in $x_1, \cdots, x_r$ such that if the numerical values $p_1, \cdots, p_r$ are substituted for the corresponding variables $x_1, \cdots, x_r$ then the resulting value is the probability that a functioning path exists from node s to node t. (If the x_i's were simply Boolean variables, this polynomial would be identical with the *structure function* of the system [5].) The reliability polynomial can be concisely expressed using two operations $\oplus$ and $\otimes$ defined on polynomials.

To begin, let

$$T^a = x_1^{a_1} x_2^{a_2} \cdots x_r^{a_r}$$

denote a monomial *term*, where each $a_i \in \{0, 1\}$. The operation $\otimes$ when applied to terms T^a and T^b yields the term T^c, where $c_i = \max\{a_i, b_i\}$. This operation is extended to arbitrary polynomials by distributivity. The operation $\oplus$ is defined on polynomials $f(\mathbf{x})$ and $g(\mathbf{x})$ using

$$f(\mathbf{x}) \oplus g(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}) - f(\mathbf{x}) \otimes g(\mathbf{x}).$$

Operations related to $\oplus$ and $\otimes$ were apparently first suggested by Mine [13] and by Kim et al. [10]. More recently Gondran and Minoux [8], and Shier [24], have formulated network reliability using the operations $\oplus$ and $\otimes$ defined above.

Let S denote the set of all polynomials that can be formed from monomial terms $T^a, T^b, \cdots$ by finite applications of the operations $\oplus$ and $\otimes$. Then it can be demonstrated [24] that $(S, \oplus, \otimes)$ forms a distributive lattice with smallest element 0 (the zero polynomial) and largest element 1 (the unit polynomial). Suppose that P_{st} is the set of simple paths from node s to node t in G. Define the *value* $v(P)$ of path P to be the product, with respect to $\otimes$, of the edge variables along the path

$$v(P) = \otimes \prod \{x_k : k \in P\}.$$

Then the reliability polynomial $F_{st}(\mathbf{x})$ can be expressed as

$$F_{st}(\mathbf{x}) = \oplus \sum \{v(P) : P \in P_{st}\}. \tag{1}$$

As an illustration, consider the standard bridge network in Fig. 1 having $s = 1$ and $t = 4$. Since there are four simple paths extending from s to t, equation (1) becomes

$$F_{st}(\mathbf{x}) = x_1 x_4 \oplus x_1 x_3 x_5 \oplus x_2 x_4 x_6 \oplus x_2 x_5.$$

Equation (1) is just the standard expression for the inclusion-exclusion formula, applied to paths in the network [1]. Expanding such an expression, using the definitions of $\oplus$

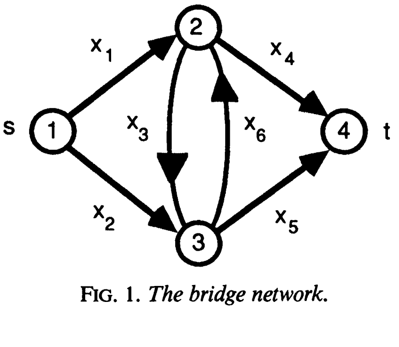


FIG. 1. *The bridge network.*

and $\otimes$, and then substituting numerical values p_k for the corresponding variables x_k yields $R_{st}(G)$.

A number of techniques have been developed to calculate the quantity (1) for a general algebraic structure (S, $\oplus$, $\otimes$) satisfying appropriate properties [6], [8]. These techniques can all be viewed as different methods of solving the system of equations

$$\mathbf{z} = \mathbf{z}M \oplus \mathbf{e}_s, \tag{2}$$

which is linear in the operations $\oplus$ and $\otimes$. Here $M = (m_{ij})$ is the *weighted adjacency matrix* for G, with $m_{ij} = x_k$ for $k = (i, j) \in E$ and $m_{ij} = 0$ otherwise. Also, $\mathbf{e}_s$ denotes the sth unit row vector. An (extremal) solution $\mathbf{z}$ to these equations is known [6] to satisfy $z_j = F_{sj}(\mathbf{x})$ for all $j \in N$. Thus by solving such equations to find $F_{st}(\mathbf{x})$, and hence $R_{st}(G)$, we also obtain the (s, j)-terminal reliabilities for *all* $j \in N$. Moreover, unlike existing methods for calculating two-terminal network reliability based on paths [1], [11], such algebraic methods do not need to first enumerate all simple paths joining the two terminals [6], [8]. These paths are automatically generated in the course of solving the set of equations (2).

A natural way of solving (2) is by means of an iterative procedure, whereby the current estimate for $\mathbf{z}$ is substituted into the right-hand side of (2), producing a new estimate for the solution vector $\mathbf{z}$. In the next section, we discuss a specific iterative scheme for solving (2) that incorporates special data structures to streamline such computations.

3. An iterative scheme. The basic idea of the iterative scheme presented here is that of passing on, at each step, the information available at node i to each of its *neighbors* j, where $(i, j) \in E$. Before stating the general iterative scheme, the ideas will first be illustrated using the network in Fig. 1. We will find all reliabilities $z_j = F_{sj}(\mathbf{x})$ relative to the source node $s = 1$.

In the algorithm, a polynomial *label* is associated with each node j. At any stage, LABEL(j) will be a reliability polynomial based on a certain subset of paths from node s to node j. In this sense, LABEL(j) corresponds to a current estimate of the solution z_j to (2). Initially, if there is an edge $k = (s, j) \in E$ then LABEL(j) = x_k. If there is no such edge then LABEL(j) = 0; in the case of the source node, LABEL(s) = 1. Those nodes, apart from s, receiving a nonzero initial label are placed on a list L. In this example, we have

j:	1	2	3	4
LABEL(j):	1	x_1	x_2	0
L:	[3, 2]			

Now we remove the "top" node i from L and update its neighbors j using

$$\text{LABEL}(j) := \text{LABEL}(j) \oplus [\text{LABEL}(i) \otimes x_k], \tag{3}$$

where $k = (i, j)$. The above (i, j) *update* simply incorporates into LABEL(j) new paths from s to j that use the edge (i, j). Any node j whose label is changed by (3) is placed on L if it does not already appear. This steps removes, in our example, $i = 3$ and updates

$$\text{LABEL}(2) = x_1 + x_2x_6 - x_1x_2x_6,$$

$$\text{LABEL}(4) = x_2x_5,$$

$$L = [2, 4].$$

The corresponding network, with node labels attached, is shown in Fig. 2.

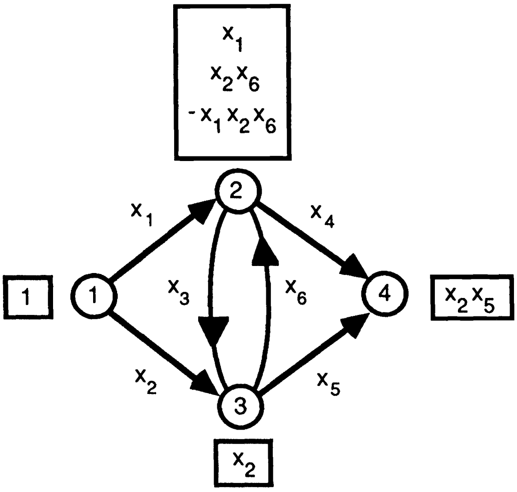


FIG. 2. *Labeling produced by first step.*

At the next step, node $i = 2$ is removed from the top of L. Nodes 3 and 4 are then updated, and node 3 is added to L:

$$\begin{aligned}\text{LABEL}(4) &= x_2x_5 + x_1x_4 - x_1x_2x_4x_5 + x_2x_4x_6 - x_2x_4x_5x_6\\ &\quad - x_1x_2x_4x_6 + x_1x_2x_4x_5x_6,\\ \text{LABEL}(3) &= x_2 + x_1x_3 - x_1x_2x_3 + x_2x_3x_6 - x_2x_3x_6\\ &\quad - x_1x_2x_3x_6 + x_1x_2x_3x_6\\ &= x_2 + x_1x_3 - x_1x_2x_3,\\ L &= [4, 3].\end{aligned}$$

This process is continued until L becomes empty. At this point, the polynomial label on any node j represents $z_j = F_{sj}(\mathbf{x})$. Table 1 shows the final labels for our example, together with the value obtained by substituting the common edge reliability p for all x_k.

The general form of the iterative procedure is specified by the following algorithm, where L again represents the list of nodes whose labels have been changed.

TABLE 1

	$F_{sj}(\mathbf{x})$	$R_{sj}(G)$, $p_k = p$
$j = 1$	1	1
$j = 2$	$x_1 + x_2x_6 - x_1x_2x_6$	$p + p^2 - p^3$
$j = 3$	$x_2 + x_1x_3 - x_1x_2x_3$	$p + p^2 - p^3$
$j = 4$	$x_2x_5 - x_1x_2x_4x_6 + x_1x_2x_4x_5x_6$ $+ x_2x_4x_6 - x_2x_4x_5x_6 + x_1x_4$ $- x_1x_2x_4x_5 + x_1x_3x_5 - x_1x_3x_4x_5$ $- x_1x_2x_3x_5 + x_1x_2x_3x_4x_5$	$2p^2 + 2p^3 - 5p^4 + 2p^5$

```
1. [Initialization]
   for $j \neq s$ do
      if $k = (s, j) \in E$ then LABEL($j$) := $x_k$
      else LABEL($j$) := 0;
   LABEL($s$) := 1;
   $L := [j: (s, j) \in E]$;
2. [Iterative Step]
      while $L \neq [\ \ ]$ do
         remove $i$ from $L$;
            for $k = (i, j) \in E$ do
               $T$ := LABEL($j$) $\oplus$ [LABEL($i$) $\otimes$ $x_k$];
               if $T \neq$ LABEL($j$) then
                  LABEL($j$) := $T$;
                  if $j \notin L$ then enter $j$ into $L$.
```

Upon termination of the algorithm, LABEL(j) will be the required reliability polynomial $F_{sj}(\mathbf{x})$. Notice that there are several ways of managing the list L. In our example, we treated L as a queue, whereby nodes are processed in a FIFO (first-in-first-out) manner. It is also possible to treat L as a stack, whereby nodes are processed in a LIFO (last-in-first-out) manner. The effect of these two ways of managing L will be examined in § 5. First, we discuss a number of useful properties of this iterative algorithm.

4. Properties. In this section we make use of the algebraic properties of $(S, \oplus, \otimes)$ to establish certain properties of the iterative algorithm presented in § 3. It will be convenient to denote the variable attached to edge (i, j) by x. Also, the label on node j at the start of step m will be denoted by $L_m(j)$. Then the (i, j) update (3) of node j after step m is expressed as

$$L_{m+1}(j) = L_m(j) \oplus xL_m(i). \tag{4}$$

Because the label on node j represents the sum with respect to $\oplus$ of a set of simple s-j paths and because this set of paths can expand through subsequent updates (4), we have

Property 1. If $k \leqq m$ then $L_k(j) \oplus L_m(j) = L_m(j)$.

One important simplification derives from the following property. It states that only the "new" information $N(i)$ added to the label of i since i was last on L needs to be propagated to its neighbors j.

Property 2. Suppose that at step m an (i, j) update is to be performed, where the labels on i and j are $L_m(i) = L_k(i) \oplus N(i)$ and $L_m(j)$ with $k < m$. Step k represents the step at which an (i, j) update previously occurred. Then at step $m + 1$ the new label assigned to j will be $L_{m+1}(j) = L_m(j) \oplus xN(i)$.

Proof. At step k, node j receives the label $L_{k+1}(j) = L_k(j) \oplus xL_k(i)$. Also, since $k < m$ we have $L_{k+1}(j) \oplus L_m(j) = L_m(j)$, by Property 1. Then

$$\begin{aligned} L_{m+1}(j) &= L_m(j) \oplus xL_m(i) \\ &= [L_m(j) \oplus L_{k+1}(j)] \oplus x[L_k(i) \oplus N(i)] \\ &= [L_m(j) \oplus L_k(j) \oplus xL_k(i)] \oplus xL_k(i) \oplus xN(i) \\ &= [L_m(j) \oplus L_k(j) \oplus xL_k(i)] \oplus xN(i) \\ &= [L_m(j) \oplus L_{k+1}(j)] \oplus xN(i) \\ &= L_m(j) \oplus xN(i). \end{aligned}$$ □

Because the labels on each node will be maintained as fully expanded polynomials (expressed using ordinary + and ×), it is desirable to know when certain of the terms in $L_m(i)$ do not affect the label $L_m(j)$. The following property provides one such condition.

Property 3. Suppose that at step m an (i, j) update is to be performed with $L_m(i) = A_1 + A_2 + \cdots + A_v$, $L_m(j) = B_1 \oplus B_2 \oplus \cdots \oplus B_w$ and $B_1 \subseteq xA_1$. Then the updated label $L_{m+1}(j) = L_m(j) \oplus x[A_2 + \cdots + A_v]$.

Proof. Let $A = A_2 + \cdots + A_v$ and $B = B_2 \oplus \cdots \oplus B_w$. Then

$$\begin{aligned} L_{m+1}(j) &= L_m(j) \oplus xL_m(i) \\ &= [B_1 \oplus B] \oplus x[A_1 + A] \\ &= [B_1 \oplus B] + [xA_1 + xA] - [xA_1 + xA][B_1 \oplus B] \\ &= [B_1 \oplus B] + xA_1 + xA - xA_1[B_1 + B - B_1B] - xA[B_1 \oplus B] \\ &= [B_1 \oplus B] + xA_1 + xA - xA_1 - xA_1B + xA_1B - xA[B_1 \oplus B] \\ &= L_m(j) + xA - xAL_m(j) \\ &= L_m(j) \oplus x[A_2 + \cdots + A_v]. \end{aligned}$$ □

Together, Properties 2 and 3 show that certain "cancellations" in the update step (4) of the iterative scheme can be predicted in advance, and thus unnecessary computation can be avoided. The next property demonstrates that the approximations to $R_{sj}(G)$, derived from successive labels at node j, are monotone nondecreasing. The notation $R_m(j)$ indicates the value obtained by substituting numerical values p_r for x_r into the polynomial $L_m(j)$.

Property 4. If $k \leqq m$ then $R_k(j) \leqq R_m(j)$.

Proof. Since $k \leqq m$ we can express $L_k(j) = T_1 \oplus T_2 \oplus \cdots \oplus T_v$ and

$$L_m(j) = T_1 \oplus T_2 \oplus \cdots \oplus T_w,$$

where T_i is a monomial term representing some path P_i from s to j and $v \leqq w$. Then $R_m(j)$ represents the probability that at least one path of $\{P_1, P_2, \cdots, P_w\}$ is functioning and so is at least as large as the probability $R_k(j)$ that at least one path of $\{P_1, P_2, \cdots, P_v\} \subseteq \{P_1, P_2, \cdots, P_w\}$ is functioning. □

5. Computational results. Several examples will be given in this section to illustrate the efficacy of a version of the iterative algorithm that makes use of Properties 2 and 3. The quality of the nondecreasing sequence of approximations to $R_{sj}(G)$ will also be examined, in particular as this relates to the discipline (FIFO, LIFO) used for managing the list L. The iterative algorithm was coded in FORTRAN 77 and all computations were performed using the IBM 3081 computer at Clemson University.

Example 1. This network, having 9 nodes and 19 edges, is taken from [20] and is shown in Fig. 3. There are 35 s-t paths and 5,287 noncancelling terms in $F_{st}(\mathbf{x})$. As discussed in Satyanarayana and Prabhakar [20], each noncancelling term corresponds to an "acyclic subgraph" of G. Despite its small size, this example represents one of the most complex directed networks whose exact reliability has been reported in the literature.

The reliabilities $R_{sj}(G)$ have been calculated using our iterative procedure and the FIFO/LIFO disciplines. For ease of presentation, the reliability polynomial $F_{st}(\mathbf{x})$ has been evaluated with all $p_k = p$ for the particular (s, t) pair indicated in Fig. 3; all edge failures are assumed to be independent. Figure 4 shows $F(p) = F_{st}(p, \cdots, p)$ plotted versus p using the FIFO discipline. As expected, the various iterations produce an increasing sequence $F1, \cdots, F9$ of reliability curves that converge to the exact answer in

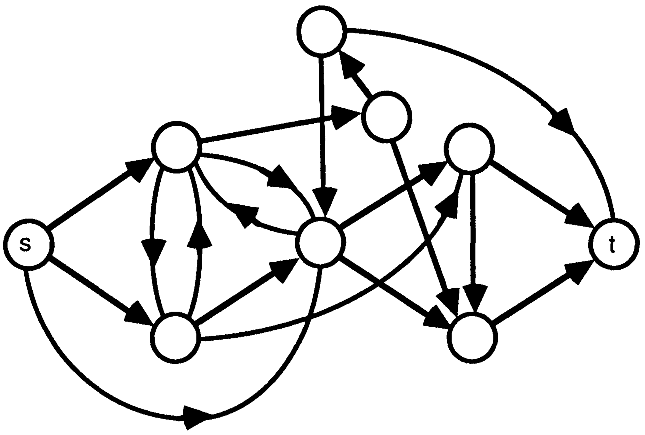


FIG. 3. *Example* 1.

9 iterations. Each iteration produces a lower bound on $R_{st}(G)$ and thus provides a *conservative* estimate for the true network reliability. Namely, the exact (s, t)-reliability of the network is guaranteed to be at least as large as the value specified by the approximation. Notice that the curves for the fifth through ninth iterations overlap in the figure, thus providing excellent approximations to $R_{st}(G)$. Also indicated in Fig. 4 are the cumulative CPU times (in seconds) required to complete the work through the end of the specified iteration. Thus, a total of 0.638 seconds were needed to obtain $R_{st}(G)$, whereas only 0.061 seconds were needed to obtain an approximation that is virtually indistinguishable over the entire range $0 \leqq p \leqq 1$.

Figure 5 shows analogous information relative to the LIFO discipline. In this case, twelve iterations were required before convergence was obtained. (Several of the curves overlap so only 10 approximations are apparent in the figure.) Although the exact answer

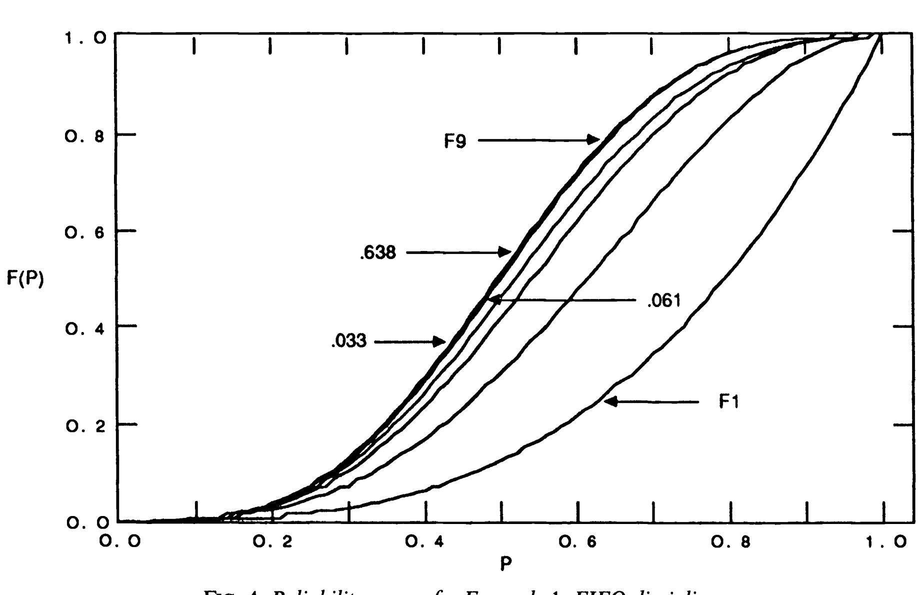


FIG. 4. *Reliability curves for Example* 1, *FIFO discipline.*

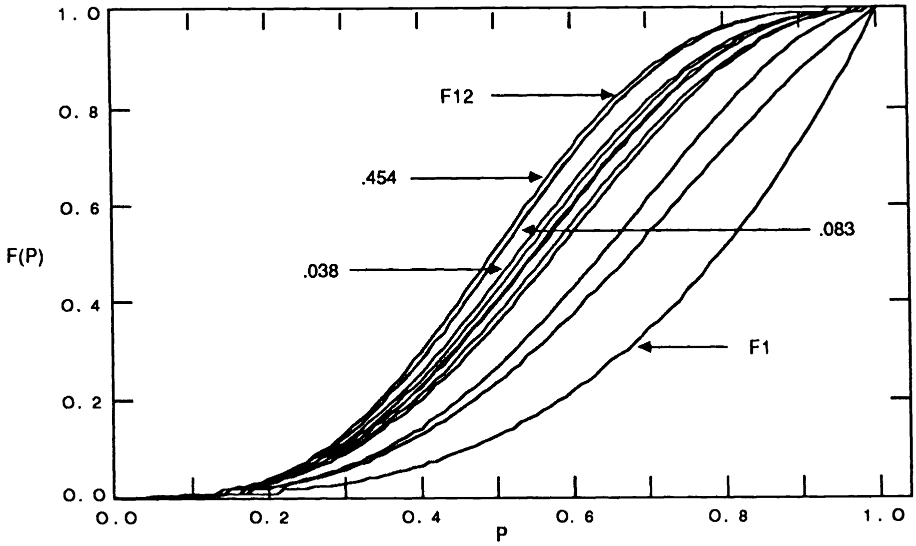


FIG. 5. *Reliability curves for Example* 1, *LIFO discipline.*

was obtained in 0.454 seconds (less than the comparable time for FIFO), the LIFO discipline did not give as useful a set of approximations compared to the FIFO approach.

Example 2. This network, with 13 nodes and 27 edges, is derived from an example given by Martelli [12]; see Fig. 6. It is considerably more complex than Example 1, having 70 s-t paths and 34,983 noncancelling terms. Plots of $F(p)$ versus p are shown in Figs. 7 and 8 for the FIFO and LIFO disciplines, respectively. Again it is observed that the LIFO method obtains the exact answer faster than the FIFO method. However, the quality of approximations produced by FIFO is superior to those produced by LIFO. Indeed, a very close approximation to the exact reliability polynomial is obtained by FIFO in 1.18 seconds, one-eighth of the time required to find the exact answer using FIFO and one-sixth of that required using LIFO.

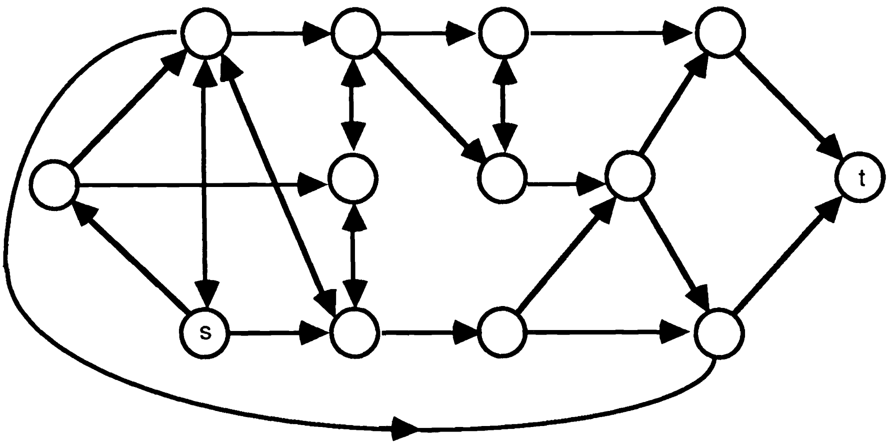


FIG. 6. *Example* 2.

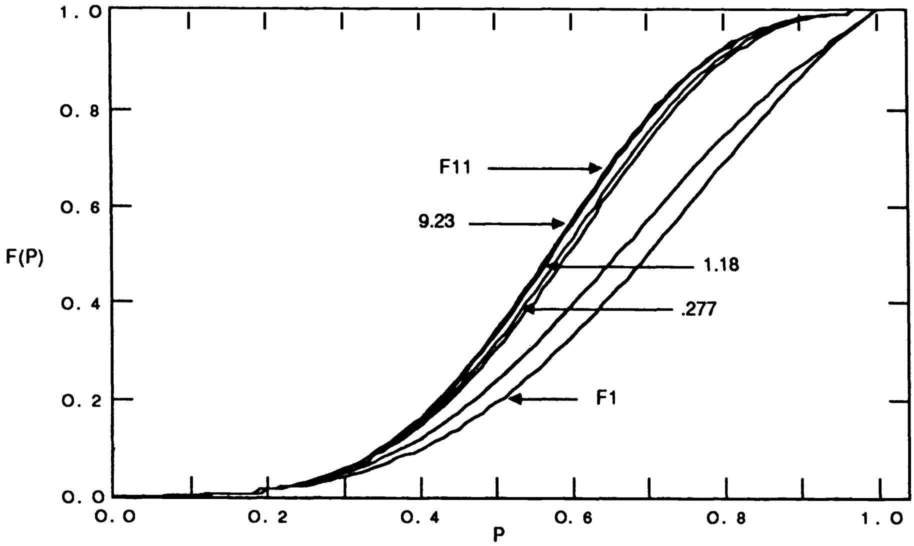


FIG. 7. *Reliability curves for Example 2, FIFO discipline.*

Finally, five random networks on 12 nodes and 30 edges were generated for test purposes. The characteristics of these networks, together with the number of iterations required for convergence, are shown in Table 2. In order to compare the quality of the approximations generated for these examples, we have tabulated the CPU time (in seconds) required to achieve a relative error of $\alpha\%$ or less (at $p = 0.5$) in Table 3. The results for Examples 1 and 2 are also included.

In these random examples the FIFO and LIFO disciplines appear to be comparable in terms of the time required to obtain the exact answer. Again, however, the FIFO

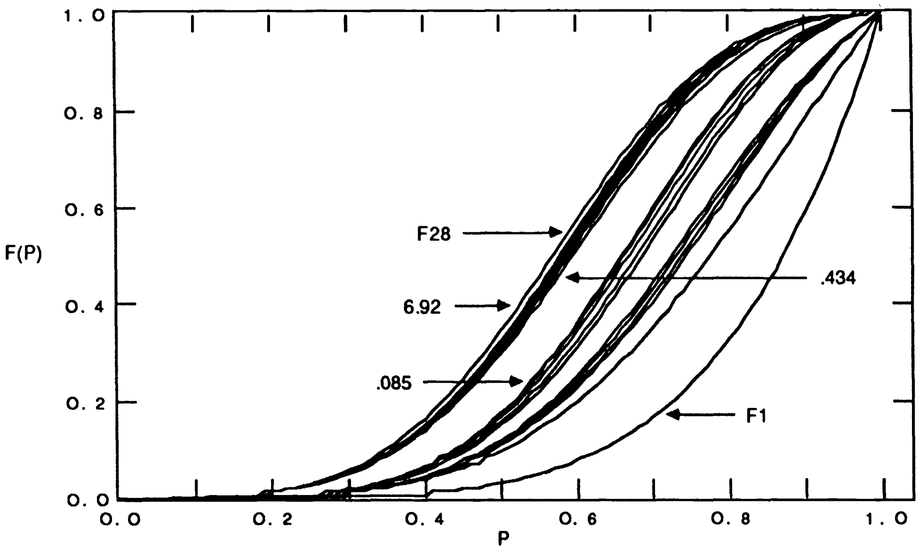


FIG. 8. *Reliability curves for Example 2, LIFO discipline.*

TABLE 2

Network	Number of s-t paths	Number of noncancelling terms	Number of iterations	
			FIFO	LIFO
R1	14	1,263	9	7
R2	28	3,383	11	8
R3	41	7,583	8	10
R4	44	17,919	5	5
R5	34	42,687	10	8

variant gives a fairly close approximation rather quickly and it completely dominates the LIFO variant in this respect.

6. Conclusions. This paper has explored an algebraic structure underlying certain network reliability problems. A promising iterative algorithm has been developed that allows both exact and approximate answers to be obtained. Rather than giving simply a single number, this algorithm produces a reliability polynomial that can then be easily evaluated at any particular input values $p_1, \cdots, p_r$. Also, in the process of determining $R_{st}(G)$ we also generate $R_{sj}(G)$ for all $j \in N$.

Empirical results have shown that the choice of data structure (FIFO, LIFO) can have a significant effect on the relative efficiency of the procedures as well as on the quality of the approximations. Whereas the LIFO approach frequently obtains the exact reliability polynomial faster than the FIFO approach, the latter produces better approximations—ones that are quite close to the exact answer but are obtained in a fraction of the time. This desirable feature of the FIFO approach can be explained as follows, assuming that the p_k are comparable in value. Under a FIFO discipline, nodes are processed in order of increasing distance from s. Thus, the first time node j is labelled, it is done so

TABLE 3

Example	Discipline	CPU (secs) for accuracy within α%			
		0%	1%	5%	10%
1	FIFO	.638	.248	.061	.033
	LIFO	.454	.454	.083	.083
2	FIFO	9.23	4.26	1.18	.277
	LIFO	6.92	6.92	6.92	.847
R1	FIFO	.049	.001	.001	.001
	LIFO	.054	.054	.004	.001
R2	FIFO	.380	.076	.006	.003
	LIFO	.362	.362	.362	.362
R3	FIFO	1.21	.172	.172	.004
	LIFO	1.11	.630	.162	.071
R4	FIFO	3.06	.195	.004	.004
	LIFO	2.85	2.85	.089	.005
R5	FIFO	5.07	.062	.002	.001
	LIFO	5.11	1.58	1.58	.002

relative to a path with the minimum number of edges. More generally, the FIFO approach ensures that the "more probable" (fewer edge) paths are incorporated as soon as possible. Subsequent (longer and less probable) paths contribute, but not as much, to the final label on node *j*. On the other hand, a LIFO discipline creates a depth-first rather than a breadth-first search of the network, and thus "early" approximations can be substantially improved by the incorporation of later (shorter) paths.

The approximate solutions generated by the iterative algorithm will always produce (conservative) lower bounds on the exact solution. If greater accuracy is required, such lower bounds can be used together with a simulation approach, such as Fishman's sampling procedure [7], that makes explicit use of lower bounds to obtain improved estimates. Alternatively, these lower bounds can be used in conjunction with existing techniques that produce upper bounds on network reliability [22], [25] to obtain an interval that must enclose $R_{st}(G)$.

Finally, it should be emphasized that regardless of the list discipline used, some relatively challenging directed networks from the literature can be solved by our algorithm with a modest amount of computation. In particular, one example studied had 70 paths. We are not aware of any existing algorithm that has exactly solved a problem of this complexity. While the proposed approach appears to have potential, further experimentation will be necessary before any firm conclusions can be drawn concerning its general applicability. In order to solve larger, more realistic problems, it may be possible to combine this approach with methods for decomposing the network into more manageable portions [23], [26].

Acknowledgments. This paper benefited greatly from the careful scrutiny of the referees.

REFERENCES

[1] A. Agrawal and R. E. Barlow, *A survey of network reliability and domination theory*, Oper. Res., 32 (1984), pp. 478–492.

[2] A. Agrawal and A. Satyanarayana, *An O(|E|) time algorithm for computing the reliability of a class of directed networks*, Oper. Res., 32 (1984), pp. 493–515.

[3] ———, *Network reliability analysis using* 2*-connected digraph reductions*, Networks, 15 (1985), pp. 239–256.

[4] M. O. Ball, *Complexity of network reliability computations*, Networks, 10 (1980), pp. 153–165.

[5] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York, 1975.

[6] B. Carré, *An algebra for network routing problems*, J. Inst. Math. Appl., 7 (1971), pp. 273–294.

[7] G. S. Fishman, *A Monte Carlo sampling plan for estimating reliability parameters and related functions*, Technical Report UNC/ORSA/TR-85/7, Univ. North Carolina, Chapel Hill, NC, 1985.

[8] M. Gondran and M. Minoux, *Graphs and Algorithms*, John Wiley, Chichester, 1984.

[9] C. L. Hwang, F. A. Tillman and M. H. Lee, *System-reliability evaluation techniques for complex/large systems—a review*, IEEE Trans. Rel., R-30 (1981), pp. 416–422.

[10] Y. Kim, K. Case and P. Ghare, *A method for computing complex system reliability*, IEEE Trans. Rel., R-21 (1972), pp. 215–219.

[11] M. Locks, *Recursive disjoint products: a review of three algorithms*, IEEE Trans. Rel., R-31 (1982), pp. 33–35.

[12] A. Martelli, *A gaussian elimination algorithm for the enumeration of cut sets in a graph*, J. Assoc. Comput. Mach., 23 (1976), pp. 58–73.

[13] H. Mine, *Reliability of physical system*, IRE Trans. Circuit Theory, CT-6 (1959), pp. 138–151.

[14] T. Politof and A. Satyanarayana, *Network reliability and inner-four-cycle-free graphs*, Math. Oper. Res., 11 (1986), pp. 484–505.

[15] ———, *A linear time algorithm to compute the reliability of planar cube-free graphs*, Technical Report, Dept. of Quantitative Methods, Concordia University, Quebec, 1985.

[16] J. S. PROVAN, *The complexity of reliability computations in planar and acyclic graphs*, Technical Report UNC/ORSA/TR-83/12, Univ. of North Carolina, Chapel Hill, NC, 1984.

[17] J. S. PROVAN AND M. O. BALL, *The complexity of counting cuts and of computing the probability that a graph is connected*, SIAM J. Comput., 12 (1983), pp. 777–788.

[18] A. ROSENTHAL, *Computing the reliability of complex networks*, SIAM J. Appl. Math., 32 (1977), pp. 384–393.

[19] A. SATYANARAYANA AND M. CHANG, *Network reliability and the factoring theorem*, Networks, 13 (1983), pp. 107–120.

[20] A. SATYANARAYANA AND A. PRABHAKAR, *A new topological formula and rapid algorithm for reliability analysis of complex networks*, IEEE Trans. Rel., R-27 (1978), pp. 82–100.

[21] A. SATYANARAYANA AND R. K. WOOD, *A linear-time algorithm for computing K-terminal reliability in series-parallel networks*, SIAM J. Comput., 14 (1985), pp. 818–832.

[22] J. G. SHANTHIKUMAR, *Simple bounds for network reliability*, Technical Report, School of Business Administration, Univ. of California, Berkeley, CA, 1984.

[23] D. R. SHIER, *A decomposition algorithm for optimality problems in tree-structured networks*, Discrete Math., 6 (1973), pp. 175–190.

[24] ———, *Iterative algorithms for calculating network reliability*, in Graph Theory with Applications to Algorithms and Computer Science, Y. Alavi et al., eds., John Wiley, New York, 1985, pp. 741–752.

[25] A. SHOGAN, *Sequential bounding of the reliability of a stochastic network*, Oper. Res., 24 (1976), pp. 1027–1044.

[26] R. E. TARJAN, *Fast algorithms for solving path problems*, J. Assoc. Comput. Mach., 28 (1981), pp. 594–614.

[27] L. G. VALIANT, *The complexity of enumeration and reliability problems*, SIAM J. Comput., 8 (1979), pp. 410–421.

[28] R. K. WOOD, *A factoring algorithm using polygon-to-chain reductions for computing K-terminal network reliability*, Networks, 15 (1985), pp. 173–190.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

012

GROUP CONVOLUTIONS AND MATRIX TRANSFORMS*

DAVID EBERLY† AND PAUL HARTUNG‡

Abstract. Given a finite group G (possibly noncommutative) and a field $\mathbb{F}$, group convolutions are constructed based on the group algebra of G over $\mathbb{F}$. Matrices with entries in the group algebra are constructed so that they have a convolution property relative to G. As special cases, the discrete Fourier transform, the discrete Walsh transform and a transform based on the dihedral groups are discussed. The development also shows that higher-dimensional transforms are special cases of the construction where the underlying group is an external direct product of other groups. An illustration of the ideas is given using the dihedral groups and matrix representations. Finally, a generalization of convolution is discussed in terms of group rings.

Key words. convolution, discrete transforms, group algebras

AMS(MOS) subject classifications. 94A11, 15A33

1. Introduction. The ideas in this paper are motivated by the concept of a discrete transform in signal processing. We briefly discuss the discrete Fourier transform and the discrete Walsh transform. Historically, both transforms are developed by considering series expansions in terms of orthogonal functions. The construction of the transforms comes about as a transition from an analytical framework to a discrete setting. From a computational point of view, it is desirable to avoid this transition and instead develop the framework of discrete transforms from a common, discrete foundation.

Such a foundation is provided based on abstract groups. As a consequence, the discrete Fourier and Walsh transforms are constructed in a more natural way than through an analytical-to-discrete approach. Multidimensional transforms can also be constructed via this foundation by looking at direct products of groups.

Extensions of these ideas have been developed in the past decade by constructing discrete transforms using surrogate number systems; for example, finite fields, Mersenne numbers, and Fermat numbers are used [1]–[3], [10]–[12], [19]–[22].

An important concept in studying discrete transforms is that of the convolution of sequences. Convolutions arise naturally in determining transfer functions (by discrete Fourier transforms), digital signal processing (using discrete Walsh transforms), correlation studies, multiplication of large integers (using number theoretic transforms), etc. Fast transforms are desirable in computing convolutions and the abundance of results in the literature support this relationship. However, the transforms used are based on Abelian groups.

In this paper, we examine the idea of convolution based on abstract groups in general (commutative or noncommutative) and matrix transforms based on these convolutions. Such transforms will have elements in a group algebra generated by a given group. An illustration of these ideas is given using the noncommutative dihedral groups.

2. The discrete Fourier transform. Let Z be the set of integers and let $\mathbb{R}$ be the set of real numbers. Define $\phi_n(x) = \exp(-inx)$, $n \in Z$, $x \in \mathbb{R}$, and $i^2 = -1$. The set $G = \{\phi_n(x): n \in Z\}$ is an orthogonal set of functions on the interval $[0, 2\pi]$ with $\int_0^{2\pi} \phi_n(x)\bar{\phi}_m(x)dx = 2\pi\delta_{nm}$. Here, δ_{nm} is the Kronecker delta and $(\bar{\ })$ is the complex

* Received by the editors April 10, 1985; accepted for publication (in revised form) September 9, 1986.

† Division of Mathematics, Computer Science, and Systems Design, University of Texas, San Antonio, Texas 78285.

‡ Department of Mathematics, Bloomsburg University, Bloomsburg, Pennsylvania 17815.

conjugate. Also, the set G forms a group under function multiplication: If ϕ_n, $\phi_m \in G$, then $\phi_n\phi_m = \phi_{n+m} \in G$ (closure); $\phi_0 = 1$ (identity); $\phi_n^{-1} = \phi_{-n}$ (inverses); the operation is clearly associative. Note that "conjugate" and "inverse" are the same operation: $\bar{\phi}_n = \phi_{-n} = \phi_n^{-1}$.

A periodic function $f(x)$ with period 2π can be represented by a Fourier series

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n\phi_n(x), \qquad c_n = \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{\phi_n}(t)dt.$$

For a given N, approximate the coefficients c_n, $n = 0, \cdots, N-1$, by the trapezoid rule to obtain

$$c_n = \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{\phi_n}(t)dt \cong \frac{1}{N}\sum_{j=0}^{N-1} f\left(\frac{2\pi j}{N}\right)\overline{\phi_n}\left(\frac{2\pi j}{N}\right) = \hat{c}_n, \qquad n=0, \cdots, N-1.$$

This system of equations is called the *discrete Fourier transform of order N* and can be written in matrix form, $M_N\mathbf{f} = \hat{\mathbf{c}}$, where $\mathbf{f} = [f_j]$ and $\hat{\mathbf{c}} = [\hat{c}_j]$ are $N \times 1$ vectors with $f_j = f(2\pi j/N)$. The matrix is given by $M_N = [\alpha^{kj}]$ where $\alpha = \exp(2\pi i/N)$. The matrix has the property $M_N^{-1} = \bar{M}_N^T/N$. For more detailed discussions on the discrete Fourier transform and properties, see [4], [13], [14], [25], [26].

3. The discrete Walsh transform. The set of functions on [0, 1] given by

$$\phi_0(x) = 1, \qquad 0 \leqq x < 1,$$

$$\phi_{1,1}(x) = \left(\begin{array}{ll} 1, & 0 \leqq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leqq x < 1 \end{array}\right),$$

$$\phi_{n+1,2k-1}(x) = \left(\begin{array}{ll} \phi_{n,k}(2x), & 0 \leqq x < \frac{1}{2} \\ (-1)^{k+1}\phi_{n,k}(2x-1), & \frac{1}{2} \leqq x < 1 \end{array}\right),$$

$$\phi_{n+1,2k}(x) = \left(\begin{array}{ll} \phi_{n,k}(2x), & 0 \leqq x < \frac{1}{2} \\ (-1)^{k}\phi_{n,k}(2x-1), & \frac{1}{2} \leqq x < 1 \end{array}\right),$$

$n = 1, 2, \cdots, k = 1, 2, \cdots, 2^{n-1}$ was first analyzed extensively by J. Walsh in 1923 [23]. These functions have been used in a number of applications in recent years. They are especially useful for computer applications since the functions take on only the values ± 1.

Let m be a nonnegative integer and let $N = 2^m$. The Walsh functions can be constructed by the following "alternating" process (and the functions have at most N discontinuities on [0, 1]). The process gives a different numbering, called the Paley ordering, than that used originally. The reason for constructing the functions in this way is that the functions are essentially listed from smallest number to largest number of discontinuities.

Partition [0, 1] into N subintervals of equal length and label these

$$I_j = [j/N, (j+1)/N], \qquad j = 0, \cdots, N-1.$$

Since the Walsh functions take on only the values ± 1, we need to specify on each of the intervals I_j what the function value will be. The argument will be illustrated with the case $N = 8$. See Fig. 1. The column numbers refer to the index of I_j; the row numbers give the new numbering for the Walsh functions, say $\psi_j(x)$.

$j=$	0	1	2	3	4	5	6	7
0	1	1	1	1	1	1	1	1
1	1	1	1	1	−1	−1	−1	−1
2	1	1	−1	−1	1	1	−1	−1
3	1	1	−1	−1	−1	−1	1	1
4	1	−1	1	−1	1	−1	1	−1
5	1	−1	1	−1	−1	1	−1	1
6	1	−1	−1	1	1	−1	−1	1
7	1	−1	−1	1	−1	1	1	−1

FIG. 1

The value for the Walsh function $\psi_j(x)$ on [0, 1/8] is always 1. For the second interval [1/8, 2/8], half of the functions have value 1 while the other half have value -1. This is shown in Fig. 1 by dividing column 1 into halves and either "duplicating" (multiplying by 1) or "complementing" (multiplying by -1) the associated entries in column 0. We have now constructed 2 columns of the matrix. To construct the next 2 columns, we divide these columns into fourths and alternately duplicate and complement the associated entries in the first 2 columns. We have now constructed 4 columns of the matrix. To construct the next 4 columns, we divide these columns into eighths and alternately duplicate and complement the associated entries which appear in the first 4 columns. This matrix represents the values of the first 8 Walsh functions (with at most 8 points of discontinuity on [0, 1]) on the subintervals I_j. Such a matrix is an example of a Hadamard matrix (a matrix such that $M^{-1} = cM^T$ for some constant c). A discussion of these can be found in [5].

Call the $N \times N$ matrix constructed by this process W_N. The set of functions $\{\psi_0, \cdots, \psi_{N-1}\}$ is isomorphic to Z_2^m, where $Z_2 = \{0, 1\}$ is the cyclic group with addition modulo 2 ($1 + 1 = 0$). To see this, in the construction of ψ_j (row j of W_N), assume first 2^k ($0 \leqq k < m$) entries of row j have been computed. To compute the next 2^k entries, we multiply the first 2^k entries by 1 or -1. If the multiplication is by 1, assign to this operation the element $0 \in Z_2$. If the multiplication is by -1, assign to this operation the element $1 \in Z_2$. There will be m such assignments in the construction of ψ_j. If the list of assignments is treated as a number in base 2, then this number is equal to the row number j in base 10.

For example, in Fig. 1, $\psi_3 \sim 0 \times 1 \times 1$ and $\psi_6 \sim 1 \times 1 \times 0$ ($3_{10} = 011_2$ and $6_{10} = 110_2$). We can see that this process, when applied for the value m, gives us a set isomorphic to Z_2^m. When the process is applied for the value $m + 1$, the set obtained is isomorphic to $Z_2 \times Z_2^m = Z_2^{m+1}$. As a consequence, the ordering of the functions does not change from that of the previous step.

Let $G = \{\psi_n(x)\}_{n=0}^{\infty}$ be the sequence of functions constructed in the above process. The set G is a complete set of orthogonal functions on [0, 1] with $\int_0^1 \psi_n(x)\overline{\psi_m}(x)dx = \delta_{nm}$. This set also forms a group under function multiplication: If $\psi_n, \psi_m \in G$, then $\psi_n\psi_m \in G$ (closure—this follows from the isomorphism property of subsets of G discussed above); $\psi_0 = 1$ (identity); $\psi_n^{-1} = \psi_n$ (inverses); the operation is clearly associative. This is an identical formulation to the one given for Fourier series. Again, note that "conjugate" and "inverse" are the same operation: $\overline{\psi_n} = \psi_n = \psi_n^{-1}$.

We can in fact compute the result $\psi_n\psi_m$ by using the isomorphism property of subsets of G. Define $n \oplus m$ to be addition of the numbers n and m in base 2 with no carries. Then $\psi_n\psi_m = \psi_{n\oplus m}$. For example, $5 \oplus 7 = 101 + 111 = 010 = 2$ (no carries used), so $\psi_5\psi_7 = \psi_2$.

A periodic function $f(x)$ with period 1 can be represented by a Walsh series

$$f(x) \sim \sum_{n=0}^{\infty} c_n \psi_n(x), \qquad c_n = \int_0^1 f(t)\overline{\psi_n}(t)dt.$$

For a given value $N = 2^m$, approximate the coefficients c_n, $n = 0, \cdots, N-1$, by the trapezoid rule to obtain

$$c_n = \int_0^1 f(t)\overline{\psi_n}(t)dt \cong \frac{1}{N}\sum_{j=0}^{N-1} f\left(\frac{j}{N}\right)\psi_n\left(\frac{j}{N}\right) = \hat{c}_n, \qquad n = 0, \cdots, N-1.$$

This system of equations is called the *discrete Walsh transform of order N* and can be written in matrix form, $W_N \mathbf{f} = \hat{\mathbf{c}}$, where $\mathbf{f} = [f_j]$ and $\hat{\mathbf{c}} = [\hat{c}_j]$ are $N \times 1$ vectors with $f_j = f(j/N)$. The matrix W_N is the matrix which was constructed earlier and has the property $W_N^{-1} = W_N/N$. For more detailed discussions on Walsh transforms, see [6]–[9], [15]–[18].

4. Group convolutions. Let G be a multiplicative group with index set I, say $G = \{g_i : i \in I\}$, where I is a subset of the integers Z. Let $i, j, k \in I$ and $g_i, g_j, g_k \in G$. Define *index summation*, $\oplus$, by: $i \oplus j = k$ iff $g_i g_j = g_k$.

Suppose G is a finite group of order N. Select the index set to be $I = \{0, \cdots, N-1\}$ and require $g_0 = 1$ (the identity element of G). Let $\mathbb{F}$ be any field and let $\mathbf{e}_i \in \mathbb{F}^N$ $(i \in I)$ be the vector which has all components 0 except for the ith component which is 1 (and the components are numbered 0 through $N-1$). If G is an infinite group, select $I = Z$ and require $g_0 = 1$. Let $\mathbb{F}^\omega$ be the set of all sequences $\mathbf{x} = \{x_i\}_{i=-\infty}^{\infty}$, where $\mathbf{x}: Z \to \mathbb{F}$. Let $\mathbf{e}_i \in \mathbb{F}^\omega$ $(i \in Z)$ be the sequence which has all components 0 except for the ith component which is 1.

Let $\mathbf{u} = \sum_{i \in I} u_i \mathbf{e}_i$ and $\mathbf{v} = \sum_{j \in I} v_j \mathbf{e}_j$ be elements in $\mathbb{F}^N$ (or $\mathbb{F}^\omega$). Define the *group convolution* of $\mathbf{u}$ and $\mathbf{v}$ to be

$$\mathbf{u} * \mathbf{v} = \sum_{k \in I}\left(\sum_{i \oplus j = k} u_i v_j\right)\mathbf{e}_k.$$

The group convolution is bilinear and since $\mathbf{e}_n = \sum_{i \in I} \delta_{ni}\mathbf{e}_i$, we have the property

$$\mathbf{e}_n * \mathbf{e}_m = \sum_{k \in I}\left(\sum_{i \oplus j = k} \delta_{ni}\delta_{mj}\right)\mathbf{e}_k = \mathbf{e}_{n \oplus m}.$$

Note that if G is not a commutative group, then $\mathbf{u} * \mathbf{v} \neq \mathbf{v} * \mathbf{u}$.

The *linear convolution* arises in this development by allowing G to be isomorphic to Z, say $G = \{x^i : i \in Z\}$. This group is multiplicative where $g_i = x^i$, $g_0 = 1$, and index summation is simply addition of integers: $g_i g_j = x^i x^j = x^{i+j} = g_{i+j}$. The sequences used in the linear convolution, $\{u_i\}_{i=-\infty}^{\infty}$, have the property $u_i = 0$ for $i < 0$ and $i > p$ (for some positive exponent p). For such sequences, say $\{u_i\}_{i=0}^{p}$ and $\{v_j\}_{j=0}^{p}$, the linear convolution lists the coefficients of the product of two polynomials:

$$\left(\sum_{i=0}^{p} u_i x^i\right)\left(\sum_{j=0}^{p} v_j x^j\right) = \sum_{k=0}^{p} (\mathbf{u} * \mathbf{v})_k x^k.$$

The *circular convolution* arises by allowing $G = \{x^i : i \in Z_N\}$, a multiplicative group isomorphic to Z_N, the cyclic group of order N. Index summation is addition modulo N. For example, if $N = 4$, then the circular convolution of $\mathbf{u}, \mathbf{v} \in \mathbb{F}^4$ is

$$(\mathbf{u}*\mathbf{v})_0=u_0v_0+u_1v_3+u_2v_2+u_3v_1, \qquad (\mathbf{u}*\mathbf{v})_2=u_0v_2+u_1v_1+u_2v_0+u_3v_3,$$

$$(\mathbf{u}*\mathbf{v})_1=u_0v_1+u_1v_0+u_2v_3+u_3v_2, \qquad (\mathbf{u}*\mathbf{v})_3=u_0v_3+u_1v_2+u_2v_1+u_3v_0.$$

The circular convolutions lists the coefficients of a product of polynomials in one variable with restriction $x^N = 1$.

The *Walsh convolution* arises by allowing G to be isomorphic to Z_2^m, a product of m copies of Z_2. The group G is multiplicative and has 2^m elements, $g_0 = 1, g_1, \cdots, g_{N-1}$ $(N = 2^m)$; each element is its own inverse. The index summation $i \oplus j = k$ can be computed as in § 3: Convert i, j to base 2 and add using no carries. For example, if $m = 2$ $(N = 4)$, then the Walsh convolution of $\mathbf{u}, \mathbf{v} \in \mathbb{F}^4$ is

$$(\mathbf{u}*\mathbf{v})_0=u_0v_0+u_1v_1+u_2v_2+u_3v_3, \qquad (\mathbf{u}*\mathbf{v})_2=u_0v_2+u_1v_3+u_2v_0+u_3v_1,$$

$$(\mathbf{u}*\mathbf{v})_1=u_0v_1+u_1v_0+u_2v_3+u_3v_2, \qquad (\mathbf{u}*\mathbf{v})_3=u_0v_3+u_1v_2+u_2v_1+u_3v_0.$$

The Walsh convolution lists the coefficients of a product of polynomials in m variables, $x_1, \cdots, x_m$, with the restrictions $x_j^2 = 1$ $(1 \leqq j \leqq m)$. The reason for naming the Walsh convolution is partly due to the fact that copies of Z_2 appear (compare with the development for the Walsh functions). A discussion on applications of the convolution can be found in [1], [10], [12]–[14].

As a last example, let $G = D_3 = \{1, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$, the dihedral group (noncommutative) with rotation element σ (of order 3) and reflection element τ (of order 2). With the selection $g_0 = 1$, $g_1 = \sigma$, $g_2 = \sigma^2$, $g_3 = \tau$, $g_4 = \tau\sigma$, $g_5 = \tau\sigma^2$, and the properties $\sigma^3 = 1$, $\tau^2 = 1$, and $\tau\sigma = \sigma^2\tau$, the table listing index summation is given in Fig. 2 below.

⊕	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	0	5	3	4
2	2	0	1	4	5	3
3	3	4	5	0	1	2
4	4	5	3	2	0	1
5	5	3	4	1	2	0

FIG. 2

The group convolution of $\mathbf{u}, \mathbf{v} \in \mathbb{F}^6$ is

$$(\mathbf{u}*\mathbf{v})_0=u_0v_0+u_1v_2+u_2v_1+u_3v_3+u_4v_4+u_5v_5,$$

$$(\mathbf{u}*\mathbf{v})_1=u_0v_1+u_1v_0+u_2v_2+u_3v_4+u_4v_5+u_5v_3,$$

$$(\mathbf{u}*\mathbf{v})_2=u_0v_2+u_1v_1+u_2v_0+u_3v_5+u_4v_3+u_5v_4,$$

$$(\mathbf{u}*\mathbf{v})_3=u_0v_3+u_1v_4+u_2v_5+u_3v_0+u_4v_2+u_5v_1,$$

$$(\mathbf{u}*\mathbf{v})_4=u_0v_4+u_1v_5+u_2v_3+u_3v_1+u_4v_0+u_5v_2,$$

$$(\mathbf{u}*\mathbf{v})_5=u_0v_5+u_1v_3+u_2v_4+u_3v_2+u_4v_1+u_5v_0.$$

5. Matrix transforms. Let G be a multiplicative group with index set I (a subset of Z). Define the *group algebra of G over* $\mathbb{F}$ to be the set of formal sums $\mathbb{F}(G) = \{\sum_{i\in I} r_i g_i : g_i \in G,\ r_i \in \mathbb{F}$, all but a finite number of r_i are zero$\}$. This set is a ring with addition defined by

$$\sum_{i\in I} u_i g_i + \sum_{i\in I} v_i g_i = \sum_{i\in I} (u_i + v_i) g_i$$

and multiplication defined by

$$\left(\sum_{i\in I} u_i g_i\right)\left(\sum_{j\in I} v_j g_j\right) = \sum_{k\in I}\left(\sum_{i\oplus j = k} u_i v_j\right) g_k = \sum_{k\in I} (\mathbf{u} * \mathbf{v})_k g_k.$$

We can define another group algebra of G over $\mathbb{F}$ to be the set of formal sums $\hat{\mathbb{F}}(G) = \{\sum_{i\in I} g_i r_i : g_i \in G,\ r_i \in \mathbb{F}$, all but a finite number of r_i are zero$\}$. This set is a ring which is isomorphic to $\mathbb{F}(G)$ and has similarly defined addition and multiplication operations. In § 4, the fact that convolutions list the coefficients of products of polynomials is just an observation that such products are essentially products of elements in $\mathbb{F}(G)$ (or $\hat{\mathbb{F}}(G)$) via the isomorphism of G with the appropriate set of polynomials. We will consider now only groups of finite order N and discuss matrices with elements in $\mathbb{F}(G)$.

Define $\mathfrak{M} = \{[a_{ij}] : a_{ij} \in \mathbb{F}(G),\ 0 \leqq i \leqq N-1,\ 0 \leqq j \leqq N-1\}$. Then $\mathfrak{M}$ is a ring with addition defined by

$$[a_{ij}] + [b_{ij}] = [a_{ij} + b_{ij}]$$

and multiplication defined by

$$[a_{ij}][b_{ij}] = \left[\sum_{k=0}^{N-1} a_{ik} b_{kj}\right].$$

$\mathfrak{M}$ is an associative algebra over $\mathbb{F}$ where scalar multiplication is defined by

$$c[a_{ij}] = [c a_{ij}], \qquad c \in \mathbb{F}.$$

Let $V = [\hat{\mathbb{F}}(G)]^N = \{(v_0, \cdots, v_{N-1}) : v_j \in \hat{\mathbb{F}}(G)\}$ be an N-dimensional vector space over $\mathbb{R}$ with the usual vector addition and scalar multiplication. An element $M \in \mathfrak{M}$ can be thought of as a linear operator, $M: V \to V$. Let $\mathbf{x} \in V$ be given by $\mathbf{x} = [x_i]$ and let $M = [a_{ij}]$. Define $M\mathbf{x}$ by

$$M\mathbf{x} = \sum_{n=0}^{N-1} g_n \left(\sum_{k\oplus l = n} \sum_{j=0}^{N-1} c_{ijk} x_{jl}\right)$$

where $x_j = \sum_{l=0}^{N-1} g_l x_{jl}$ and $a_{ij} = \sum_{k=0}^{N-1} c_{ijk} g_k$; $x_{jl}, c_{ijk} \in \mathbb{F}$. The definition for the operation of M on elements is motivated by the usual matrix arithmetic, where the elements of M are in a field and V is Euclidean N-space:

$$\begin{aligned}
M\mathbf{x} = \sum_{j=0}^{N-1} a_{ij} x_j &= \sum_{j=0}^{N-1} \left(\sum_{k=0}^{N-1} c_{ijk} g_k\right)\left(\sum_{l=0}^{N-1} g_l x_{jl}\right) \\
&= \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} c_{ijk} g_k g_l x_{jl} \\
&= \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} g_k g_l c_{ijk} x_{jl} \quad \text{(the definition)} \\
&= \sum_{j=0}^{N-1} \sum_{n=0}^{N-1} g_n \left(\sum_{k\oplus l = n} c_{ijk} x_{jl}\right).
\end{aligned}$$

Interchanging summations gives us the definition for $M\mathbf{x}$.

Let $\mathbf{x} = [x_i]$, $\mathbf{y} = [y_j] \in V$. Define the binary operation, $\circ$, to be the bilinear operation having the property: $\mathbf{x} \circ \mathbf{y} = [x_iy_i]$. If $M = [a_{ij}] \in \mathfrak{M}$, define M to have the *convolution property* if

$$M(\mathbf{u} * \mathbf{v}) = (M\mathbf{u}) \circ (M\mathbf{v}) \quad \text{for all } \mathbf{u}, \mathbf{v} \in \mathbb{F}^N.$$

THEOREM 1. *If $M = [a_{ij}] \in \mathfrak{M}$ has the convolution property, and if $a_{ij} = \sum_{k=0}^{N-1} c_{ijk}g_k$, then* (1) $a_{im}a_{in} = a_{i,m\oplus n}$ *and* (2) $c_{i,m\oplus n,p} = \sum_{k\oplus l=p} c_{imk}c_{inl}$.

Proof. Let $\mathbf{e}_i$ be the $N \times 1$ vector which has all components 0 except for the ith component which is 1. Then the ith component of $M\mathbf{e}_t$ is given by

$$(M\mathbf{e}_t)_i = \sum_{j=0}^{N-1} a_{ij}\delta_{jt} = a_{it}.$$

This implies that $(M\mathbf{e}_m) \circ (M\mathbf{e}_n) = [a_{im}a_{in}]$. But

$$\mathbf{e}_m * \mathbf{e}_n = \mathbf{e}_{m\oplus n}, \text{so } M(\mathbf{e}_m * \mathbf{e}_n) = M(\mathbf{e}_{m\oplus n}) = [a_{i,m\oplus n}].$$

Since M has the convolution property, $M(\mathbf{e}_m * \mathbf{e}_n) = (M\mathbf{e}_m) \circ (M\mathbf{e}_n)$. Consequently, $a_{im}a_{in} = a_{i,m\oplus n}$ and (1) is proved.

To see that (2) is also true, $a_{it} = \sum_{k=0}^{N-1} c_{itk}g_k$ and condition (1) imply that

$$\begin{aligned}\sum_{p=0}^{N-1} c_{i,m\oplus n,p}g_p &= \left(\sum_{k=0}^{N-1} c_{imk}g_k\right)\left(\sum_{l=0}^{N-1} c_{inl}g_l\right)\\ &= \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} c_{imk}c_{inl}g_kg_l\\ &= \sum_{p=0}^{N-1}\left(\sum_{k\oplus l=p} c_{imk}c_{inl}\right)g_p.\end{aligned}$$

Equating coefficients of the g_p terms gives the result in (2). □

6. Examples. The examples listed illustrate the linear convolution, the circular convolution, the Walsh convolution, and the convolution for the noncommutative group D_3.

Consider the linear convolution $(\mathbf{u} * \mathbf{v})_k = \sum_{i+j=k} u_iv_j$, $i, j \in Z$, where the sequences for $\mathbf{u}$ and $\mathbf{v}$ have only a finite number of nonzero terms. If $F(\mathbf{f}) = \sum_{n=-\infty}^{\infty} f_ne^{inx}$, then F satisfies the convolution property (discussed in the usual analysis of Fourier series) $F(\mathbf{u} * \mathbf{v}) = F(\mathbf{u})F(\mathbf{v})$. This is part of the motivation for looking at convolution properties for the general convolution (derived from a group G).

Consider the circular convolution $(\mathbf{u} * \mathbf{v})_k = \sum_{i\oplus j=k} u_iv_j$ where the underlying group is Z_N, the cyclic group of order N with addition modulo N. Let $M = [\alpha^{kj}]$ be an $N \times N$ matrix where $\alpha = \exp(2\pi i/N)$. Then M has the convolution property $M(\mathbf{u} * \mathbf{v}) = (M\mathbf{u}) \circ (M\mathbf{v})$. As pointed out earlier, this matrix is the one used for the discrete Fourier transform and has the property that $M^{-1} = \bar{M}^T/N$. Compare with the introductory remarks on the group properties of $\{\exp(-inx): n \in Z\}$: If $M = [\alpha^{kj}]$, then $\bar{M} = [\alpha^{-kj}]$; that is $M = [a_{ij}]$ implies that $\bar{M} = [a_{ij}^{-1}]$.

Consider the Walsh convolution $(\mathbf{u} * \mathbf{v})$ where the underlying group is Z_2^m and index summation is computed by adding i and j base 2 with no carries. Let $W = [w_{ij}]$ be the $N \times N$ matrix constructed in § 3. Then W has the convolution property $W(\mathbf{u} * \mathbf{v}) = (W\mathbf{u}) \circ (W\mathbf{v})$. The matrix was the one used for the discrete Walsh transform and has the property that $W^{-1} = W/N = \bar{W}^T/N$. If $W = [w_{ij}]$, then $\bar{W} = [w_{ij}^{-1}]$.

Finally, consider the dihedral group D_3. The convolution was given in § 4 along with the table for index summation. Listed in Fig. 3 is a matrix $S = [s_{ij}]$ which has the convolution property with respect to D_3. The block matrix $R = [\sigma^{ij}]$ and the block matrix $\tau R = [\tau\sigma^{ij}]$, where $0 \leqq i \leqq 2$ and $0 \leqq j \leqq 2$.

$$S = \begin{pmatrix} R & -\tau R \\ R & \tau R \end{pmatrix}$$

FIG. 3

It is easy to verify that the entries of S satisfy the condition $s_{im}s_{in} = s_{i,m\oplus n}$. As in the previous examples, S has the property $S^{-1} = \bar{S}^T/N$ and $\bar{S} = [s_{ij}^{-1}]$ where $(\bar{\ })$ is complex conjugation acting on the elements of D_3; that is, $\bar{\sigma} = \sigma^2$, and $\bar{\tau} = \tau$.

7. Convolutions and transforms for direct products. We have seen in the examples that the discrete Fourier transform of order N arises as a matrix which has the convolution property relative to the group Z_N. We have also seen that the discrete Walsh transform of order N arises as a matrix which has the convolution property relative to the group Z_2^m, $N = 2^m$. Groups provide a common origin for both types of discrete transforms. In addition, the development here actually allows the construction of transforms which represent the higher dimensional discrete Fourier and Walsh transforms. For a good discussion, see [27].

As a prelude, we make the observation that the group convolution depends on the numbering of the elements of G. This is important when considering index summation. Superficially, this may seem to be a burden to the development. For example, if $\alpha = \exp(2\pi i/6)$ and $\beta = \exp(2\pi i/3)$, then Z_6 is isomorphic to $\{1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5\}$ and $Z_2 \times Z_3$ is isomorphic to the group $\{(1, 1), (1, \beta), (1, \beta^2), (-1, 1), (-1, \beta), (-1, \beta^2)\}$ where multiplication is performed componentwise. If $\mathbf{u}, \mathbf{v} \in \mathbb{F}^6$, then relative to Z_6,

$$(\mathbf{u} * \mathbf{v})_0 = u_0v_0 + u_1v_5 + u_2v_4 + u_3v_3 + u_4v_2 + u_5v_1$$

and relative to $Z_2 \times Z_3$

$$(\mathbf{u} * \mathbf{v})_0 = u_0v_0 + u_1v_2 + u_2v_1 + u_3v_3 + u_4v_5 + u_5v_4.$$

Clearly the group convolutions differ even though Z_6 and $Z_2 \times Z_3$ are isomorphic groups. It turns out that this can be quite useful when considering higher dimensional transforms. However, the labeling of vectors will be done differently.

Let $G = \{g_0, \cdots, g_{N-1}\}$ and $H = \{h_0, \cdots, h_{M-1}\}$ be finite groups where $g_0 = 1$ (the identity in G) and $h_0 = 1$ (the identity in H). Let $\oplus$ be the index summation for G and let $\oplus'$ be the index summation for H. The external direct product of G and H is $G \times H = \{p_{ij} = (g_i, h_j): 0 \leqq i \leqq N-1, 0 \leqq j \leqq M-1\}$. List the elements of $G \times H$ so that the subscripts increase (where ij is treated as a two-digit number in base max (N, M)). This set is a group with product

$$p_{ij}p_{kl} = (g_i, h_j)(g_k, h_l) = (g_ig_k, h_jh_l) = (g_{i\oplus k}, h_{j\oplus' l}) = p_{i\oplus k, j\oplus' l}.$$

The order of $G \times H$ is NM. Let $[u_{ij}]$ and $[v_{ij}]$ be $N \times M$ arrays with elements in $\mathbb{F}$. Define the group convolution of $\mathbf{u} = [u_{ij}]$ and $\mathbf{v} = [v_{ij}]$ relative to $G \times H$ by

$$(\mathbf{u} * \mathbf{v})_{nm} = \sum_{i\oplus k = n} \sum_{j\oplus' l = m} u_{ij}v_{kl}, \qquad 0 \leqq n \leqq N-1, \qquad 0 \leqq m \leqq M-1.$$

The motivation, as before, is that the product of two elements in the group algebra $\mathbb{F}(G \times H)$ looks like

$$\left(\sum_{i=0}^{N-1}\sum_{j=0}^{M-1} u_{ij}p_{ij}\right)\left(\sum_{k=0}^{N-1}\sum_{l=0}^{M-1} v_{kl}p_{kl}\right) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} (\mathbf{u}*\mathbf{v})_{nm}p_{nm}.$$

Note also that the convolution defined here for $G \times H$ is different than that for $H \times G$.

Let $\mathbf{e}_{rs} = \mathbf{e}_r \otimes \mathbf{e}_s = [\delta_{ir}\delta_{js}]$, the matrix with all entries 0 except for the rth row and sth column entry which is 1. Here, $\otimes$ is the tensor product. The definition for group convolution implies that $\mathbf{e}_{rs}*\mathbf{e}_{tu} = \mathbf{e}_{r\oplus t, s\oplus' u}$.

Let $V = \{[x_{ij}]: x_{ij} \in \hat{\mathbb{F}}(G\times H)\}$. Then V is a vector space under the usual matrix addition and scalar multiplication (with scalars in $\mathbb{F}$). We wish to discuss matrices whose entries are in $\mathbb{F}(G \times H)$ and such that as an operator, we have $M: V \rightarrow V$. To do so, we need the following. Let $A = [a_{ik}]$ be an $I \times K$ matrix and let $B = [b_{jl}]$ be a $J \times L$ matrix where a_{ik} and b_{jl} are elements of a ring. Define the *Kronecker product* $A \times B$ to be the $IJ \times KL$ matrix whose entry in row $iJ + j$ and column $kL + l$ is given by $c_{ij,kl} = a_{ik}b_{jl}$, $0 \leqq i \leqq I - 1, 0 \leqq j \leqq J - 1, 0 \leqq k \leqq K - 1, 0 \leqq l \leqq L - 1$. $A \times B$ can be interpreted as an $I \times K$ array of $J \times L$ blocks with the (i, k)th block given by $a_{ik}B$. Note that $A \times B \neq B \times A$ and $(A \times B) \times C = A \times (B \times C)$. It is also true that $(A \times B)(C \times D) = (AC) \times (BD)$ whenever the matrix multiplications AC and BD are defined.

For example, if $A = [\alpha^{ij}]$, $\alpha = -1$, and $B = [\beta^{ij}]$, $\beta = \exp(2\pi i/3)$, then

$$A \times B = \begin{pmatrix} B & B \\ B & \alpha B \end{pmatrix} \quad \text{and} \quad B \times A = \begin{pmatrix} A & A & A \\ A & \beta A & \beta^2 A \\ A & \beta^2 A & \beta A \end{pmatrix},$$

both 6×6 arrays.

We can construct matrices with entries in $\mathbb{F}(G \times H)$ by the following process. Let $A = [a_{ik}]$ be an $N \times N$ matrix with entries in $\mathbb{F}(G)$ and let $B = [b_{jl}]$ be an $M \times M$ matrix with entries in $\mathbb{F}(H)$ where $N = |G|$, and $M = |H|$. Identify the element a_{ik} with $(a_{ik}, 1) \in \mathbb{F}(G \times H)$ and the element b_{jl} with $(1, b_{jl}) \in \mathbb{F}(G \times H)$. Then $A \times B$ is the $NM \times NM$ matrix whose entry in row $iM + j$ and column $kM + l$ is

$$(a_{ik}, 1)(1, b_{jl}) = (a_{ik}, b_{jl}) \in \mathbb{F}(G\times H),$$

$0 \leqq i \leqq N - 1, 0 \leqq k \leqq N - 1$, and $0 \leqq j \leqq M - 1, 0 \leqq l \leqq M - 1$. Thus, A, B, and $A \times B \in \mathfrak{M} = \{[c_{ij}]: c_{ij} \in \mathbb{F}(G \times H)\}$.

Given $\mathbf{x} = [x_{ij}] \in V$, define $T: V \rightarrow V$ by

$$(T\mathbf{x})_{nm} = \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} (a_{ni}, b_{mj})x_{ij}, \qquad x_{ij} \in \hat{\mathbb{F}}(G\times H).$$

If $\mathbf{u} = [u_{ij}]$ and $\mathbf{v} = [v_{ij}]$ are elements in V, then define the binary operation $\circ$ by $\mathbf{u} \circ \mathbf{v} = [u_{ij}v_{ij}]$. Define the matrix T to have the *convolution property* relative to $G \times H$ if

$$T(\mathbf{u}*\mathbf{v}) = (T\mathbf{u}) \circ (T\mathbf{v}), \qquad \mathbf{u}, \mathbf{v} \in V.$$

We have the following result.

THEOREM 2. *If $A \in \mathfrak{M}(\mathbb{F}(G))$ has the convolution property relative to G and if $B \in \mathfrak{M}(\mathbb{F}(H))$ has the convolution property relative to H, then $A \times B \in \mathfrak{M}(\mathbb{F}(G \times H))$ has the convolution property relative to $G \times H$.*

Proof. It is sufficient to consider $A \times B$ applied to the basis tensors $\mathbf{e}_{ij}$. From the definition of matrix multiplication for $A \times B$, we have that

$$(A\times B)\mathbf{e}_{rs} = \sum_{i=0}^{N-1}\sum_{j=0}^{M-1} (a_{ni}, b_{mj})\delta_{ir}\delta_{js} = [(a_{nr}, b_{ms})]$$

using the δ-substitution property; similarly, $(A \times B)\mathbf{e}_{tu} = [(a_{nt}, b_{mu})]$. We also have

$$\begin{aligned}(A \times B)(\mathbf{e}_{rs} * \mathbf{e}_{tu}) &= (A \times B)(\mathbf{e}_{r \oplus t, s \oplus' u}) = [(a_{n, r\oplus t}, b_{m, s \oplus' u})] \\ &= [(a_{nr}a_{nt}, b_{ms}b_{mu})] = [(A \times B)\mathbf{e}_{rs}] \circ [(A \times B)\mathbf{e}_{tu}],\end{aligned}$$

where Theorem 1 has been used. □

The ideas in this section clearly generalize to products of the form

$$\prod_{k=1}^{n} G_k = \{(g_{i_1}^{(1)}, \cdots, g_{i_n}^{(n)}) : g_{i_k}^{(k)} \in G_k, 0 \leqq i_k \leqq N_k - 1 = |G_k| - 1\}.$$

If A_k is a matrix with entries in $\mathbb{F}(G_j)$ and has the convolution property relative to G_j, then $A_1 \times \cdots \times A_k \in \mathbb{F}(G_1 \times \cdots \times G_k)$ has the convolution property relative to $G_1 \times \cdots \times G_k$. This matrix has the property: if $M = A_1 \times \cdots \times A_k$, then

$$M = [a_{1ij}] \times \cdots \times [a_{kij}], \qquad \bar{M} = [a_{1ij}^{-1}] \times \cdots \times [a_{kij}^{-1}],$$

and $M^{-1} = c\bar{M}^T$, where $c = [|G_1| \cdots |G_k|]^{-1}$; here we use $A_n^{-1} = c_n \bar{A}_n^T$, $A_n = [a_{nij}]$, and $\bar{A} = [a_{nij}^{-1}]$.

8. Examples. Using the development for direct products of groups, we can illustrate higher dimensional transforms.

The two-dimensional discrete Fourier transform can be constructed by considering groups of the type $Z_N \times Z_M$. If $\alpha = \exp(2\pi i/N)$ and $\beta = \exp(2\pi i/M)$, then $A = [\alpha^{ij}]$ has the convolution property relative to Z_N, $B = [\beta^{ij}]$ has the convolution property relative to Z_M, and $A \times B = [\alpha^{ij}\beta^{kl}]$ has the convolution property relative to $Z_{N \times M}$. The two-dimensional discrete Fourier transform is given by

$$\hat{c}_{nm} = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \alpha^{ni}\beta^{mj} f_{ij}$$

where f_{ij} is a sampled $N \times M$ array of data.

The two-dimensional discrete Walsh transform can be constructed by considering groups of the type $Z_2^{m_1} \times Z_2^{m_2}$. If $N_1 = 2^{m_1}$ and $N_2 = 2^{m_2}$, then the matrices W_{N_1} and W_{N_2} are constructed as shown in § 3. Each has the appropriate convolution property. The matrix $W_{N_1} \times W_{N_2}$ has the convolution property relative to the full group product. Note that, however, $W_{N_1} \times W_{N_2} \neq W_{N_1 N_2}$. For example,

$$W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad W_2 \times W_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad \text{and}$$

$$W_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Of course, the three- and higher-dimensional transforms can be constructed by using three or more groups in the direct product formulation.

9. An illustration using dihedral groups. The main result in this paper is the idea of developing convolution and matrix transforms from the group algebra point of view with the thought of developing applications using noncommutative groups. To construct matrix transforms with the convolution property, one needs to construct inverses of elements in the group algebra. In general, this is a difficult process. To avoid this difficulty, one can resort to matrix representations of the group algebra elements. The idea is illustrated using the dihedral groups.

Let G be the dihedral group with generators τ ($\tau^2 = 1$) and σ ($\sigma^N = 1$). Let $g_i = \sigma^{i/2}$ if i is even and $g_i = \tau\sigma^{(i-1)/2}$ if i is odd ($0 \leqq i \leqq N-1$). Then $g_ig_j = g_{i\oplus j}$ defines the index summation $i \oplus j = j + (-1)^j i + \delta N$ where $\delta \in Z$ is chosen so that the right-hand side of the equation is between 0 and $N-1$.

Let $\mathbb{F}$ be any field and consider the dihedral group algebra $\mathbb{F}(G)$. If $\alpha \in \mathbb{F}(G)$, then

$$\alpha = \sum_{i=0}^{N-1} a_{2i}\sigma^i + \sum_{i=0}^{N-1} a_{2i+1}\tau\sigma^i.$$

Define the group algebra element $\hat{\alpha}$ by

$$\hat{\alpha} = \sum_{j=0}^{N-1} a_{2j}\sigma^{-j} - \sum_{j=0}^{N-1} a_{2j+1}\tau\sigma^j.$$

Then

$$\alpha\hat{\alpha} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} (a_{2i}a_{2j} - a_{2i+1}a_{2j+1})\sigma^{i-j} = \sum_{p=-(N-1)}^{N-1} \sum_{i-j=p} (a_{2i}a_{2j} - a_{2i+1}a_{2j+1})\sigma^p$$

for $0 \leqq i \leqq N-1$, $0 \leqq j \leqq N-1$. For the summation over p, the positive indexed terms give

$$\sum_{i-j=p} (a_{2i}a_{2j} - a_{2i+1}a_{2j+1})\sigma^p = \sum_{j=0}^{N-1-p} [a_{2(j+p)}a_{2j} - a_{2(j+p)+1}a_{2j+1}].$$

The negative indexed terms give

$$\sum_{i-j=-p} (a_{2i}a_{2j} - a_{2i+1}a_{2j+1})\sigma^{-p} = \sum_{j=p}^{N-1} [a_{2(j-p)}a_{2j} - a_{2(j-p)+1}a_{2j+1}].$$

The $p = 0$ term gives

$$\sum_{i-j=0} (a_{2i}a_{2j} - a_{2i+1}a_{2j+1}) = \sum_{i=0}^{N-1} (a_{2i}^2 - a_{2i+1}^2).$$

It is easily shown that the summations for $i - j = p$ and $i - j = -p$ are equal. Call this common value c_p. Then

$$\alpha\hat{\alpha} = c_0 + \sum_{p=1}^{N-1} c_p(\sigma^p + \sigma^{-p}).$$

We can identify the elements of G with operations on vectors in $\mathbb{F}^2$. Let the matrix representation for τ be T and let the matrix representation for σ be S. Represent the identity 1 by the identity matrix I. Then

$$T = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad S = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}, \quad \theta = \frac{2\pi}{N}.$$

Let $\mathfrak{D}$ be the set of matrix representations of the group algebra $\mathbb{F}(G)$ with the above identifications. For $\alpha \in \mathbb{F}(G)$, let $M(\alpha) \in \mathfrak{D}$ be its representation. Let $\alpha = \sum a_ig_i$, where c_p are defined as before. Then $M(\alpha)$ is invertible if and only if $\frac{1}{2}c_0 + \sum c_p \cos(p\theta) \neq 0$. This follows from $M(\alpha\hat{\alpha}) = M(\alpha)M(\hat{\alpha})$ and the formula for the product $\alpha\hat{\alpha}$:

$$M(\alpha\hat{\alpha})=c_0M(1)+\sum_{p=1}^{N-1}c_p[M(\sigma^p)+M(\sigma^{-p})]=\left[c_0+2\sum_{p=1}^{N-1}c_p\cos(p\theta)\right]I$$

so that $M(\alpha\hat{\alpha})=kI$. If $k\neq 0$, then $M(\alpha)$ and $M(\hat{\alpha})$ are invertible and $M(\alpha)^{-1}=M(\hat{\alpha})/k$. Note the striking similarity to the properties for the discrete Fourier and Walsh transforms: $M^{-1}=\bar{M}^T/N$ ($\hat{}$ corresponds to $\bar{}$). Of course, it is not necessarily true that in $\mathbb{F}(G)$, α and $\hat{\alpha}/k$ are inverses. But in the matrix transforms with elements from $\mathbb{F}(G)$, instead of finding inverses for the entries, we can substitute the values $\hat{\alpha}/k$ and resort to computations with matrix representations. Figure 4 shows the application of matrix representations to a matrix which has the convolution property. Here $G=\{1,\sigma,\sigma^2,\tau,\tau\sigma,\tau\sigma^2\}$ (compare with § 6).

$$S=\begin{bmatrix}R & -\tau R\\ R & \tau R\end{bmatrix}=\begin{bmatrix}1&1&1&-\tau&-\tau&-\tau\\ 1&\sigma&\sigma^2&-\tau&-\tau\sigma&-\tau\sigma^2\\ 1&\sigma^2&\sigma&-\tau&-\tau\sigma^2&-\tau\sigma\\ 1&1&1&\tau&\tau&\tau\\ 1&\sigma&\sigma^2&\tau&\tau\sigma&\tau\sigma^2\\ 1&\sigma^2&\sigma&\tau&\tau\sigma^2&\tau\sigma\end{bmatrix}$$

$$M(S)=\left[\begin{array}{cc|cc|cc|cc|cc|cc}
1&0&1&0&1&0&-1&0&-1&0&-1&0\\
0&1&0&1&0&1&0&1&0&1&0&1\\ \hline
1&0&a&-b&a&b&-1&0&-a&b&-a&-b\\
0&1&b&a&-b&a&0&1&b&a&-b&a\\ \hline
1&0&a&b&a&-b&-1&0&-a&-b&-a&b\\
0&1&-b&a&b&a&0&1&-b&a&b&a\\ \hline
1&0&1&0&1&0&1&0&1&0&1&0\\
0&1&0&1&0&1&0&-1&0&-1&0&-1\\ \hline
1&0&a&-b&a&b&1&0&a&-b&a&b\\
0&1&b&a&-b&a&0&-1&-b&-a&b&-a\\ \hline
1&0&a&b&a&-b&1&0&a&b&a&-b\\
0&1&-b&a&b&a&0&-1&b&-a&-b&-a
\end{array}\right]$$

FIG. 4

The parameters are $a=\cos(2\pi/3)$ and $b=\sin(2\pi/3)$. The matrix $M(S)$ can be used in applications in the standard ways that the discrete Fourier and discrete Walsh transforms are used.

10. Conclusions. The preceding formulation seems to be a good foundation for looking at matrix transforms which have a convolution property. There may be useful applications for transforms derived from the noncommutative groups. Groups which have elements of order 2 would be extremely useful since these elements have matrix representations which "act" like the real numbers 1 and -1, a property which makes the Walsh transform useful for digital signal processing. The quaternions might also be a group to consider since there are elements such that $g^2=1$ and $h^2=-1$.

One last note: One could generalize the concept of convolution in the following way. Recall that if G is a multiplicative group and R is a ring, then the group ring of G over R is the set of all formal sums

$$R(G)=\left\{\sum_{i\in I}r_ig_i : r_i\in R, g_i\in G, \text{ all but a finite number of } r_i \text{ are } 0\right\}.$$

The addition and multiplication are as given for the group algebras discussed earlier. The multiplication makes sense if the r_i are not "related" to the g_i. We could let $R = \mathbb{F}(G)$. If so, then allowing for interaction between R and G, we could define ring multiplication for $R(G)$ by

$$\begin{aligned}
\left(\sum_{i\in I} u_i g_i\right)\left(\sum_{j\in I} v_j g_j\right) &= \sum_{i\in I}\sum_{j\in I} u_i g_i v_j g_j, \qquad u_i, v_j \in R(G),\\
&= \sum_{i\in I}\sum_{j\in I}\left(\sum_{k\in I} u_{ik} g_k\right) g_i \left(\sum_{l\in I} v_{jl} g_l\right) g_j, \qquad u_{ik}, v_{jl} \in \mathbb{F},\\
&= \sum_{i\in I}\sum_{j\in I}\sum_{k\in I}\sum_{l\in I} u_{ik} v_{jl} g_k g_i g_l g_j\\
&= \sum_{i\in I}\sum_{j\in I}\sum_{k\in I}\sum_{l\in I} u_{ik} v_{jl} g_{k\oplus i} g_{l\oplus j}\\
&= \sum_{n\in I}\sum_{m\in I}\sum_{k\oplus i=n}\sum_{l\oplus j=m} u_{ik} v_{jl} g_n g_m\\
&= \sum_{p\in I}\left(\sum_{n\oplus m=p}\sum_{k\oplus i=n}\sum_{l\oplus j=m} u_{ik} v_{jl}\right) g_p.
\end{aligned}$$

Thus, for $\mathbf{u} = [u_i]$ and $\mathbf{v} = [v_j]$ as $N \times 1$ vectors in $[\hat{\mathbb{F}}(G)]^N$, define the group convolutions of these vectors by

$$(\mathbf{u} * \mathbf{v})_p = \sum_{n\oplus m=p}\sum_{k\oplus i=n}\sum_{l\oplus j=m} u_{ik} v_{jl}.$$

This definition implies that $(\mathbf{e}_i g_i) * (\mathbf{e}_k g_l) = \mathbf{e}_{i\oplus k} g_{j\oplus l}$. We can define a matrix M to have the convolution property if $M(\mathbf{u} * \mathbf{v}) = (M\mathbf{u}) \circ (M\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in [\hat{\mathbb{F}}(G)]^N$. As in Theorem 1, we can prove that if $M = [a_{ij}]$, then

$$a_{in} g_k a_{im} g_k^{-1} = a_{i,n\oplus m}$$

for all $g_k \in G$. If $k = 0$ ($g_0 = 1$), then $a_{in} a_{im} = a_{i,n\oplus m}$. In addition, if $n = 0$, then $a_{i0} a_{im} = a_{i,0\oplus m} = a_{im}$; if $m = 0$, then $a_{in} a_{i0} = a_{i,n\oplus 0} = a_{in}$. Thus, $a_{i0} = 1$ (the identity in the group ring).

If k is arbitrary, and since $a_{i0} = 1$, we have that $a_{i0} g_k a_{im} g_k^{-1} = a_{i,0\oplus m}$, or $g_k a_{im} = a_{im} g_k$ for all i, m, k. Thus, a_{im} must commute with every element in G. As a consequence, a_{im} can only be a linear combination of the elements of the center of G.

For an Abelian group G, the center of G is exactly G. Thus, this more general convolution allows construction of matrices which have the convolution property and the matrices can be chosen just like the ones given earlier. For noncommutative groups, this definition of convolution is very restrictive. For example, in D_3, the center consists of only the identity. No matrices (which are invertible) can be constructed which have the convolution property relative to D_3.

REFERENCES

[1] R. C. Agarwal and C. S. Burris, *Fast convolution using Fermat number transforms with applications to digital filtering*, IEEE Trans. Acoust. Speech Signal Process., ASSP-22 (1974), pp. 87–97.

[2] ———, *Number theoretic transforms to implement fast digital convolution*, Proc. IEEE, 63 (1975), pp. 550–560.

[3] D. E. Bachman and T. A. Kriz, *Computational efficiency of number theoretic transform implemented finite impulse response filters*, Electron. Lett., 14 (1978), pp. 240–246.

[4] J. W. Cooley and J. W. Tukey, *An algorithm for the machine computation of complex Fourier series*, Math. Comp., 19 (1965), pp. 297–301.

[5] R. B. Crittenden, *On Hadamard matrices and complete orthonormal systems*, Proc. 1973 Walsh Functions Symposium, National Technical Information Service, U.S. Dept. Commerce, Springfield, VA, 1973, pp. 82–84.

[6] E. Gibbs and M. J. Millard, *Walsh functions as solutions of a logical differential equation*, DES Report 1, National Physics Laboratory (1969).

[7] ———, *Some methods of linear ordinary logical differential equations*, DES Report 2, National Physics Laboratory (1969).

[8] E. Gibbs, *Some properties of functions on the nonnegative integers less than 2^n*, DES Report 3, National Physics Laboratory (1969).

[9] ———, *Functions that are solutions of a logical differential equation*, DES Report 4, National Physics Laboratory (1970).

[10] W. K. Jenkins, *Composite number theoretic transforms for digital filtering*, Proc. 9th Asilomar Conf. Circuits, Systems, Comput., (1972), pp. 421–425.

[11] H. J. Nussbaumer, *Complex convolutions via Fermat number transforms*, IBM J. Res. Develop., 20 (1976), pp. 282–284.

[12] ———, *Digital filtering using complex Mersenne transforms*, IBM J. Res. Develop., 20 (1976), pp. 498–504.

[13] H. J. Nussbaumer and P. Quandelle, *Computational of convolutions and discrete Fourier transforms*, IBM J. Res. Develop., 22 (1978), pp. 134–144.

[14] ———, *Fast computation of discrete Fourier transforms using polynomial transforms*, IEEE Trans. Acoust. Speech Signal Process., ASSP-27 (1979), pp. 169–181.

[15] R. E. Paley and N. Wiener, *Characters of Abelian groups*, Proc. Nat. Acad. Sci. USA., 19 (1933), pp. 253–257.

[16] F. Pichler, *On the state-space description of linear dyadic systems*, Proc. 1971 Walsh Functions Symposium, National Technical Information Service, U.S. Department of Commerce, Springfield, VA.

[17] ———, *Walsh functions and optimal linear systems*, Proc. 1970 Walsh Functions Symposium, National Technical Information Service, U.S. Department of Commerce, Springfield, VA, pp. 17–22.

[18] ———, *Walsh functions and linear systems*, Proc. 1970 Walsh Functions Symposium, National Technical Information Service, U.S. Department of Commerce, Springfield, VA, pp. 175–182.

[19] J. M. Pollard, *The fast Fourier transform in a finite field*, Math. Comp., 25 (1971), pp. 365–374.

[20] I. S. Reed and T. K. Truong, *The use of finite fields to compute convolutions*, IEEE Trans. Inform. Theory, IT-21 (1975), pp. 208–213.

[21] B. Rice, *Some good fields and rings for computing number theoretic transforms*, IEEE Trans. Acoust. Speech Signal Process., ASSP-27 (1979), pp. 432–433.

[22] M. C. Vanwormhoudt, *On number theoretic Fourier transforms in residue class rings*, IEEE Trans. Acoust. Speech Signal Process., ASSP-25 (1977), pp. 585–586.

[23] J. L. Walsh, *A closed set of orthogonal functions*, Amer. J. Math., 55 (1923), pp. 5–24.

[24] ———, *Walsh functions and Hadamard matrices*, Proc. 1970 Walsh Functions Symposium, National Technical Information Service, U.S. Department of Commerce, Springfield, VA, pp. 163–165.

[25] S. Winograd, *On computing the discrete Fourier transform*, Proc. Nat. Acad. Sci. USA, 73 (1976), pp. 1005–1006.

[26] ———, *On computing the discrete Fourier transform*, Math. Comp., 32 (1978), pp. 175–199.

[27] ———, *A new method for computing the discrete Fourier transform*, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., (1977), pp. 366–368.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

013

COMPLEXITY OF FINDING EMBEDDINGS IN A k-TREE*

STEFAN ARNBORG†, DEREK G. CORNEIL‡ AND ANDRZEJ PROSKUROWSKI§

Abstract. A k-tree is a graph that can be reduced to the k-complete graph by a sequence of removals of a degree k vertex with completely connected neighbors. We address the problem of determining whether a graph is a partial graph of a k-tree. This problem is motivated by the existence of polynomial time algorithms for many combinatorial problems on graphs when the graph is constrained to be a partial k-tree for fixed k. These algorithms have practical applications in areas such as reliability, concurrent broadcasting and evaluation of queries in a relational database system. We determine the complexity status of two problems related to finding the smallest number k such that a given graph is a partial k-tree. First, the corresponding decision problem is NP-complete. Second, for a fixed (predetermined) value of k, we present an algorithm with polynomially bounded (but exponential in k) worst case time complexity. Previously, this problem had only been solved for $k = 1, 2, 3$.

Key words. graph theory, k-trees, algorithm complexity, NP-complete

AMS(MOS) subject classifications. 05C10, 68C25, 90C39

1. Motivation.

The class of *k-trees* is defined recursively as follows (see, for instance, Rose [15]). The complete graph with k vertices is a k-tree. A k-tree with $n + 1$ vertices ($n \geqq k$) can be constructed from a k-tree with n vertices by adding a vertex adjacent to all vertices of one of its k-vertex complete subgraphs, and only to these vertices. In a given construction of a k-tree, the original k-complete subgraph is its *basis*. Any k-complete subgraph of a k-tree can be its basis (Proskurowski [13, Prop. 1.3]). A *partial k-tree* is a subgraph of a k-tree.

Our interest in the class of k-trees and their subgraphs is motivated by some practical questions about reliability of communication networks in the presence of constrained line- and site-failures (Farley [8], Farley and Proskurowski [9], Neufeldt and Colbourn [12], Wald and Colbourn [16]), concurrent broadcasting in a common medium network (Colbourn and Proskurowski [6]), reliability evaluation in complex systems (Arnborg [1]), and evaluation of queries in relational data base systems; for a survey see Arnborg [2]. For these problems restricted to partial k-trees, there exist efficient solution algorithms which exploit the following *separation property* of k-trees (Rose [15]): every minimal separator of a k-tree consists of k completely connected vertices. This property, together with the requirement that a graph is connected and does not contain a set of $k + 2$ completely connected vertices, is a definitional property of k-trees (Rose [15, Thm. 1.1]).

Partial k-trees have a similar *bounded decomposability property* (cf. Arnborg and Proskurowski [3]): a sufficiently large partial k-tree can be disconnected by removal of at most k (separator) vertices so that each of the resulting connected components augmented by the completely connected separator vertices is a partial k-tree. Since the component partial k-trees of a partial k-tree interact only through the minimal separators of

* Received by the editors May 31, 1984; accepted for publication (in revised form) October 7, 1986.

† Department of Numerical Analysis and Computing Science, The Royal Institute of Technology, S-100 44 Stockholm, Sweden. The work of this author was supported in part by a grant from the Swedish Board of Technical Development, STU contract 83-3719.

‡ Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4. The work of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Canada-France Science Exchange Program.

§ Computer and Information Science Department, University of Oregon, Eugene, Oregon 97403. The work of this author was supported in part by the U.S.-Sweden Scientific Cooperation Program of the National Science Foundation through grant INT-8318441.

the k-tree, solutions to subproblems on the component partial k-trees may be combined to form a solution for the given partial k-tree. Using a succinct representation of a bounded number of optimal solutions to such subproblems (cf. Corneil and Keil [7]), Arnborg and Proskurowski [4] develop a general algorithm paradigm to solve efficiently many difficult problems on graphs with this bounded decomposability property. The algorithms presented there solve NP-hard optimization problems (like vertex cover, chromatic number, graph reliability) for partial k-trees given with a suitable embedding k-tree. However, the presentation skirts two major problems: one is finding an embedding, a k-tree of which the given graph is a subgraph; the other is finding the minimal value of such k, which is important since the algorithms, though linear in the size of the input, are exponential or even superexponential in k. For small values of k, recognition and embedding problems for partial k-trees have been solved by reducing a given graph according to a complete set of safe reduction rules, i.e., application of the reduction rules (in any order) reduces a graph to the empty graph iff it is a partial k-tree (see Wald and Colbourn [16] for $k = 2$, and Arnborg and Proskurowski [3] for $k = 3$). In this paper, we first address the question of determining the minimum value of k for which a given graph is a partial k-tree. We show that the decision version of this PARTIAL K-TREE problem is NP-complete. We then follow a "brute-force" approach to finding a k-tree embedding of the given graph through examination of all its k-vertex separators. This yields a polynomial time algorithm, assuming a fixed value of k. Not surprisingly, this new algorithm is inferior to the reduction rules algorithms for $k = 2$ and $k = 3$.

2. Definitions. A graph G with vertex set V and edge set E will be denoted $G(V, E)$. The cardinality of the vertex set will be called the *size* of G. A *partial graph* of G contains all its vertices and a subset of its edges, whereas a *subgraph* of G has a subset of both edges and vertices of G. A *supergraph* of G is any graph of which G is a partial graph. For general graph theoretical concepts, the reader is advised to consult a standard text, e.g., Bondy and Murty [5].

A *clique* in a graph $G(V, E)$ is a maximal complete subgraph of G. The *clique number* $\omega(G)$ is the size of a largest clique in G. For any vertex $v \in V$, the (*open*) *neighborhood* of v is defined as the set of all vertices adjacent to v, $\Gamma(v) = \{u : (u, v) \in E\}$. The *closed neighborhood* of v contains also the vertex v. The *degree* of v is the size of its neighborhood, deg $(v) = |\Gamma(v)|$, and $\Delta(G) = \max_{v \in V} \deg(v)$. A vertex v is *simplicial* if the subgraph of G induced by $\Gamma(v)$ is complete. A graph G is *k-decomposable* iff either G has $k + 1$ or fewer vertices or there is a subgraph S of G with at most k vertices such that $G - S$ is disconnected, and each of the connected components of $G - S$ augmented by S with completely connected vertices is k-decomposable. A graph is *chordal* (or *triangulated*) if every cycle of length greater than three has a chord. Clearly, k-trees are examples of chordal graphs. An *elimination scheme* of a graph is an ordering π of its vertices. The *filled graph* of $G(V, E)$ w.r.t. π is the graph $G(V, E \cup F^{\pi})$, where F^{π} are the fill edges. An edge (u, w) is a *fill edge* if there is a vertex v preceding u and w in π such that both u and w are adjacent to v via original or fill edges but not to each other. The complete set of fill edges is easily obtained by examining vertices in order π. A graph G has a *perfect elimination scheme*, i.e., an elimination scheme with no fill edges (cf. Rose [14]), if there exists an order of eliminating the vertices of G such that each vertex is simplicial at the time of elimination. It is well known (Rose [14]) that a graph is chordal iff it has a perfect elimination scheme, and that every edge-minimal chordal supergraph of a graph G is the filled graph of G w.r.t. some elimination scheme. Given A, a complete subgraph of graph G, we say that G is an *A-chordal path* if there exists a perfect elimination scheme π such that if u immediately follows v in π then u and v are adjacent, and the vertices in A are last in π. A chordal graph G is a *chordal path* iff there is an A for which G is an A-chordal path.

A *k-chordal graph* is a chordal graph G for which $\omega(G) = k + 1$. Thus, in every perfect elimination scheme of a k-chordal graph, the neighborhood of every vertex, when eliminated, induces K_i, a complete graph with i vertices, $i \leqq k$. We notice that a k-tree with more than k vertices is a k-chordal graph. Since the neighborhood of a simplicial vertex u in a k-chordal graph is a completely connected set of at most k vertices that separate u from the rest of the graph, any k-chordal graph is k-decomposable. Given a graph G, we define $k_t(G)$ to be the minimum k such that G is a partial k-tree. Similarly, $k_c(G)$ is defined to be the minimum k such that G is a partial k-chordal graph. Not surprisingly, we have the following lemma relating $k_t(G)$ and $k_c(G)$ for any graph G.

LEMMA 2.1. *For any graph G that is not a complete graph, $k_t(G) = k_c(G)$.*

Proof. From the definitions, we see that $k_c(G) \leqq k_t(G)$. To show that $k_t(G) \leqq k_c(G)$, we let G' be a $k_c(G)$-chordal supergraph of G. Since G' is $k_c(G)$-decomposable, it follows from Arnborg and Proskurowski [3, Thm. 2.7] that G' is also a partial $k_c(G)$-tree, and so is G. □

A *block* of a graph G is a maximal set of vertices with the same closed neighborhood. Clearly, the blocks of G partition V. A *block-contiguous* elimination scheme is one in which the vertices of each block are eliminated contiguously.

Yannakakis [17] introduced the notion of chain graph: a bipartite graph $G(A \cup B, E)$ is a *chain graph* if the neighbors of the nodes in A form a chain, i.e., there exists a bijection $\tau : A \leftrightarrow \{1, 2, \cdots, |A|\}$ such that $\tau(u) < \tau(v)$ iff $\Gamma(u) \supseteq \Gamma(v)$. Such a permutation τ is called a *chain order*. The neighbors of the nodes of B also form a chain, and thus the definition is unambiguous. Given a bipartite graph $G(A \cup B, E)$ and an ordering τ of A, a *(G, τ)-chain graph* is any bipartite graph $G'(A \cup B, E \cup E')$ for which τ is a chain graph order. For a given bipartite graph $G(A \cup B, E)$, the graph $C(G)$ is formed from G by adding edges to form complete subgraphs on A and B. The following lemma relates chain graphs and chordal graphs.

LEMMA 2.2 (Yannakakis [17, Lemma 2.1]). *A bipartite graph $G(A \cup B, E)$ is a chain graph iff $C(G)$ is chordal.* □

In fact, we can strengthen this statement by a more detailed characterization of the chordal graph $C(G)$.

COROLLARY 2.3. *A bipartite graph $G(A \cup B, E)$ is a chain graph iff $C(G)$ is an A-chordal path.* □

If V is the vertex set of a graph G and τ is a permutation of V, then we define the *linear cut value of* G w.r.t. τ as

$$c_\tau(G) = \max_{1 \leqq i < |V|} |\{(u, v) \in E : \tau(u) \leqq i < \tau(v)\}|.$$

The MINIMUM CUT LINEAR ARRANGEMENT problem (MCLA) is defined as follows: Given a graph $G(V, E)$ and a positive integer k, does there exist a permutation τ of V such that $c_\tau(G) \leqq k$? MCLA is NP-complete (see, for instance, Garey and Johnson [10]). In the next section, we use this fact to show the NP-completeness of the follow--ing PARTIAL K-TREE recognition problem: Given a graph G and an integer k, is $k_t(G) \leqq k$?

3. NP-completeness of PARTIAL K-TREE. In this section we will use the concepts of chain graph and chordal path to prove that the PARTIAL K-TREE problem is at least as difficult as the MCLA problem, in the standard sense of polynomial reducibility.

Given an arbitrary graph $G(V, E)$, we will construct a bipartite graph $G'(A \cup B, E')$ in the following way: Each vertex $x \in V$ is represented by $\Delta(G) + 1$ vertices in A and

$\Delta(G) - \deg(x) + 1$ vertices in B. We let A_x (resp. B_x) denote the set of vertices in A (resp. B) which represents x. Each edge $e \in E$ is represented by two vertices in B; this set of vertices is denoted B_e. Edges in E' are of the following two types: (i) all vertices in A_x are adjacent to all vertices in B_x; (ii) all vertices in A_x are adjacent to both vertices in B_e if x is an endpoint of e. As an example of this construction, see the graph in Fig. 1. We note that the vertex sets A_v, B_v and B_e form the blocks of $C(G')$.

Before proving the main result of this section, we relate block-contiguous elimination schemes of a given graph G to chordal supergraphs of G.

LEMMA 3.1. *Let H be a minimal chordal supergraph of G. Then there exists a block-contiguous elimination order π such that H is the filled graph of G w.r.t. π.*

Proof. As stated in § 2, H is the filled graph of G w.r.t. some elimination scheme. A vertex is simplicial iff all other vertices in its block are also simplicial. Since the elimination of any vertex preserves this property, any chordal graph has a block-contiguous perfect elimination scheme. Given such a scheme any ordering of the vertices in a block determines a block-contiguous perfect elimination scheme. If u and v belong to the same block of G, then the addition of fill edge (v, w) implies the addition of fill edge (u, w). Thus the blocks of G form a refinement (possibly trivial) of the blocks of H. We now set π to be a block-contiguous perfect elimination scheme for H which is also block-contiguous for G. □

We now establish the relationship between the linear cut value of a graph G and values of k' such that $C(G')$ is a partial k'-tree.

LEMMA 3.2. *Given a graph G and a positive integer k, G has a minimum linear cut value k w.r.t. some permutation π iff the corresponding graph $C(G')$ is a partial k'-tree for $k' = (\Delta(G) + 1)(|V| + 1) + k - 1$.*

Proof. Since a k'-tree is a k'-chordal graph, it follows from Lemma 3.1 that if $C(G')$ is a partial k'-tree then there exists a block-contiguous elimination scheme π' such that no vertex has degree greater than k' when it is eliminated. Let F be the filled graph of $C(G')$ w.r.t. this permutation π'. F is also a supergraph of the filled graph of $C(G')$ w.r.t.

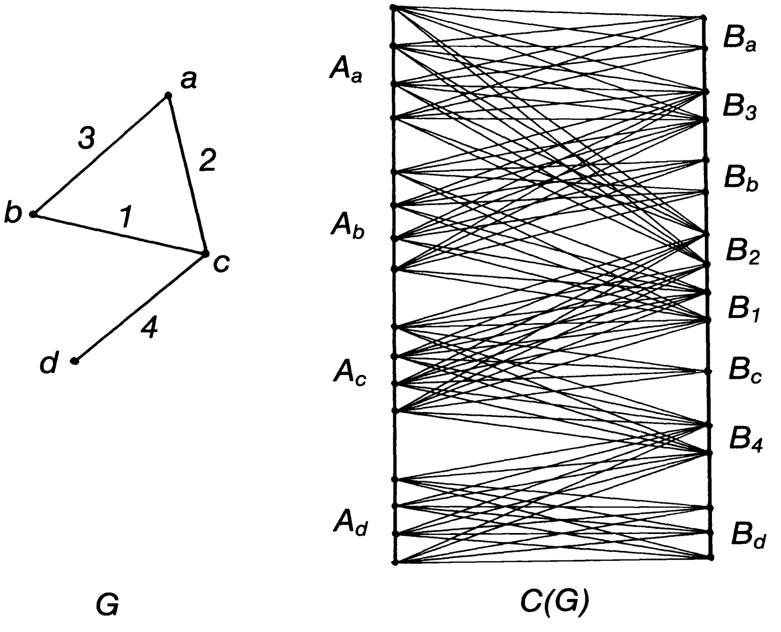


FIG. 1. *Example of the bipartite graph construction.*

any perfect elimination order of F. But since F is chordal, its edges between A and B form a chain graph by Lemma 2.2, and it is easy to see that F has a perfect elimination ordering starting with all vertices in A in reverse chain ordering, which is also contiguous in the blocks A_v. Without loss of generality we can assume that π' is such an ordering. Let π be the ordering of blocks in A induced by π'. Assume without loss of generality that the vertices of G are numbered in order π and are identified by their numbers. Consider the graph resulting from elimination of vertices in the first $i - 1$ blocks of $C(G')$. In this graph, each vertex of A_i is adjacent to: the other $\Delta(G)$ vertices in A_i; the $\Delta(G) + 1$ vertices in each of $A_{i+1}, \cdots A_{|V|}$; the $\Delta(G) + 1 - \deg_G(j)$ vertices in B_j for $j = 1, \cdots i$ (these are fill edges except for $j = i$); and the two vertices in B_e for each edge e incident to at least one vertex in $\{1, \cdots i\}$ (these are fill edges for e not incident to i). These adjacencies sum up to

$$\Delta(G) + (\Delta(G) + 1)(|V| - i) + (\Delta(G) + 1) \times i - \sum_{j=1}^{i} \deg_G(j) + 2|E_1^i| + 2|E_2^i|,$$

where E_1^i is the set of edges with exactly one vertex in $\{1, \cdots, i\}$, and E_2^i the set with both vertices in $\{1, \cdots, i\}$. Obviously, $\sum_{j=1}^{i} \deg_G(j) = 2|E_2^i| + |E_1^i|$, so the degree of a vertex in A_i simplifies to $(\Delta(G) + 1)(|V| + 1) - 1 + |E_1^i|$. Since E_1^i is the set of edges between vertices in $\{1, \cdots, i\}$ and vertices in $\{i + 1, \cdots, |V|\}$, in this particular ordering π, the maximum size of E_1^i over all i is the linear cut value of G w.r.t. π. This value also determines the maximum size of a clique in $C(G')$. We have thus shown that the k'-chordality implies the existence of a linear arrangement with the cut value k. Conversely, the existence of an ordering π w.r.t. which G has a linear cut value k implies that the largest clique in F, the filled graph of $C(G')$ w.r.t. π', has size $k' + 1$ (by examination of an induced ordering π' of its vertices.) This completes the proof. □

THEOREM 3.3. *The* PARTIAL K-TREE *problem is* NP-*complete.*

Proof. (Hardness for NP): This follows from Lemma 3.2 and the fact that $C(G')$ can be constructed from G in polynomial time. (Membership in NP): For a suitable (nondeterministic) choice of vertex order, the elimination process is easily turned into a polynomial time verification that a graph is a partial k-tree. □

Let us define a *k-chordal path* to be a k-chordal graph which is also a chordal path. A *k-interval graph* is an interval graph derived from a set of intervals, no $k + 2$ of which have a nonempty intersection (i.e., it has clique number $k + 1$ or less). We now have:

COROLLARY 3.4. *The following problems are* NP-*complete*:

(i) *Given graph G and integer k, is G a partial k-chordal path*?

(ii) *Given graph G and integer k, is G a partial k-interval graph*?

Proof. Statement (i) follows from Theorem 3.3, Lemma 2.1 and Corollary 2.3. A result of Gilmore and Hoffman [11] says that graph G is an interval graph iff the cliques of G can be numbered $C_1, C_2, \cdots, C_m$ such that for each node x, $x \in C_i \cap C_j$, $(i < j)$ implies that $x \in C_l$ for all l such that $i < l < j$. So the class of k-interval graphs is contained in the class of k-chordal graphs but contains the class of k-chordal paths, from which (ii) follows. □

4. Recognition of partial k-trees for a fixed value of k. In the preceding section, we have shown that the partial k-tree recognition problem is NP-complete if k is part of the problem's instance. Since the proof of our NP-completeness result (Theorem 3.3) builds on the value of k growing polynomially with the size of the graph, one could expect the complexity of partial k-tree recognition for fixed k to grow quickly with k. However, when the value of k is fixed (i.e., all instances of the problem refer to the same k value), the complexity status of the problem changes, as any dependence on k is considered constant. The recognition problems for partial 2- and 3-trees have been solved previously (cf. Wald and Colbourn [16] and Arnborg and Proskurowski [3]) by exhibiting complete sets of safe reduction rules, reducing to the empty graph precisely those graphs in the relevant class.

Another approach proves successful for general, fixed values of k. This new approach uses a dynamic programming technique in evaluating feasibility of proposed partial embeddings of subgraphs of the given graph in a k-tree. Although there might be many such embeddings, the set of all possible minimal separators in all embeddings has cardinality bounded by a polynomial in the size of the graph (the number of vertices), due to the fact that all such separators have size k. Our algorithm considers the connected components into which k-element vertex sets separate the graph and decides their embeddability in the order of their increasing sizes. Thus, successful embedding attempts (which assume completely connected minimal separators) can be subsequently used to embed a union of such connected components.

ALGORITHM 4.1 for the recognition of partial k-trees.

INPUT: A graph G, with n vertices.
OUTPUT: YES or NO.
DATA STRUCTURE: Family of k-element vertex sets which are separators of G. For each such set S, there is a set of l connected components of G into which G is separated by removal of S. Denoting S by C_i, we denote by C_i^j, $1 \leqq j \leqq l$ the subgraphs of G, each induced by S and the vertices of the corresponding connected component, with the addition of edges required to make the subgraph induced by S complete. Each such C_i^j has an answer YES or NO (whether it is embeddable in a k-tree or not) determined during the computation.

METHOD:

```
{find the graphs $C_i$ and $C_i^j$}
for each set $S$ of $k$ vertices in $G$ do
  if $S$ is a separator of $G$
    then insert $C_i = S$ and the corresponding graphs $C_i^j$ into the data
    structure
end-do
sort all graphs $C_i^j$ by increasing size
{examine graphs $C_i^j$ from smallest to largest and determine whether
the graph is a partial $k$-tree}
A graph $C_i^j$ of size $k + 1$ is a partial $k$-tree: set its answer to YES.
for each graph $C_i^j$ an increasing order of size $h$ do
  for each $v \in C_i^j$ do
    examine all $k$-vertex separators $C_m$ contained in $C_i \cup \{v\}$;
    consider all $C_m^l$ in $(C_i^j - C_i) \cup C_m$ which are partial $k$-trees.
    if their union, over all $l$'s and all $m$'s, contains $C_i^j - C_i$
      then set the answer for $C_i^j$ to YES and exit-do.
  end-do
  if no answer was set for $C_i^j$
    then set the answer for $C_i^j$ to NO.
  if $G$ has a separator $C_m$ such that all $C_m^l$ graphs have answer YES
    then $G$ is a partial $k$-tree: return (YES).
  if each separator $C_m$ of $G$ has a $C_m^l$ with answer NO
    then $G$ is not a partial $k$-tree: return (NO).
end-do
{end of the algorithm}
```

In the algorithm above we use C_i^j to denote both a graph and its vertex set, depending on context. This slightly inaccurate usage will continue in the verification below. The worst case time complexity of the algorithm is fairly straightforwardly bounded by a polynomial in n, the size of G, since all the operations (searches and checks) can be performed efficiently and there is a limited (polynomially bounded) number of them.

THEOREM 4.2. *The execution time of the partial k-tree recognition algorithm using suitably chosen data structures is of order* $\mathcal{O}(n^{k+2})$.

Proof. The algorithm examines at most all k-element vertex sets; there are $\mathcal{O}(n^k)$ of those, and it takes $\mathcal{O}(n^2)$ time to check if one is a separator of G. To be able to access the subgraphs C_i^j in the increasing order of their sizes, they should be bucket-sorted, in time proportional to the numer of them, at most $\mathcal{O}(n^{k+1})$. The exit conditions for the algorithm can be checked in constant time per examined subgraph, by incrementally maintaining counts of partial k-tree components for each separator, and of incorrectly guessed separators for the whole graph. There are less than n vertices in a subgraph C_i^j, and the access to a 'related' separator C_m (in the innermost loop) can be made in constant time. Checking the union of the relevant partial k-tree components is again of order of the size of C_i^j, and thus, the overall time complexity is $\mathcal{O}(n^{k+2})$. □

To prove the correctness of our algorithm we state and prove two lemmas. The first one reflects the fact that partial k-trees are k-decomposable (cf. Arnborg and Proskurowski [3, Thm. 2.7]).

LEMMA. 4.3. *A given graph G of size at least* $k + 2$ *is a partial k-tree if and only if there exists a k-vertex separator* C_i *such that all subgraphs* C_i^j *(as defined in the algorithm) are partial k-trees.*

Proof. (By induction on the size n of G). Obviously true for $n = k + 2$, since of graphs with $k + 2$ vertices only K_{k+2} is not a partial k-tree. Assuming the hypothesis true for all smaller graphs, consider a graph G of size $n \geqq k + 3$. If G is a partial k-tree, then it has a vertex v which in some k-tree embedding has a completely connected neighborhood S. The graph $G_1 = G - \{v\} \cup S$ is also a partial k-tree with a postulated separator C_i (by the inductive assumption). If this C_i is identical with S, then it fulfills the requirements for G since $S \cup \{v\}$ (the new C_i^j containing v) is a partial k-tree. Otherwise, the C_i^j (of the embedding of G_1) that contains S can be extended by v to a partial k-tree. Hence, the necessity is proved. The sufficiency follows immediately from the constructive definition of k-trees: G can be constructed using C_i as a base, and independently attaching an embedding for each C_i^j. □

The second lemma addresses the operation of the algorithm when computing the partial answers. Note that C_i^j has a complete subgraph induced by C_i.

LEMMA 4.4. *A graph* C_i^j, *as defined in the algorithm, is a partial k-tree iff there exists a vertex v in* C_i^j *and a set F of k-vertex separators* $C_m \neq C_i$ *contained in* $C_i \cup \{v\}$ *such that graphs* $C_m^l - C_m$ *(for all l such that* $C_m^l \subset C_i^j$ *is a partial k-tree) partition* $C_i^j - C_i - \{v\}$.

Proof. We recall that a k-tree can be constructed with any k-complete subgraph as its base. If C_i^j is a partial k-tree, then any k-tree T embedding it can be constructed from C_i by first adding a vertex, v, adjacent to all vertices of C_i, and then constructing the remainder of T as k-trees T_m^l based on some k-complete subgraphs of $C_i \cup \{v\}$. The k-trees T_m^l overlap only on $C_i \cup \{v\}$, and thus their subgraphs C_m^l partition $C_i^j - C_i - \{v\}$. This proves the necessity. If such a family F of separators exists, then a k-tree T embedding C_i^j can be constructed from C_i by first adding v adjacent to all the vertices of C_i and then building up the remainder of T as the union of embeddings of the partial k-trees C_m^l, each constructed with C_m as its base. □

THEOREM 4.5. *The algorithm correctly determines whether a given graph G is a partial k-tree.*

Proof. The termination criteria of the algorithm's main loop correspond to the view of partial k-trees as k-decomposable graphs. By Lemma 4.4, every subgraph C_i^j of size at most h is correctly classified. If G is a partial k-tree then the final decision (return of YES) is reached when the size of the subgraph C_i^j increases to at most $(n + k)/2$ (this corresponds to G having only k-path embeddings). □

The algorithm discussed above answers the embeddability decision problem, but it does not produce an embedding when one exists. It is obvious that this can be achieved by storing an embedding for every C_i^j if and when it has been classified as a partial k-tree.

Note added in proof. In a recent paper (*Graph minors* XIII: *The disjoint path problem*, manuscript, September 1986), Robertson and Seymour show—nonconstructively—the existence of an $\mathcal{O}(n^2)$ algorithm for recognizing partial k-trees. Such an algorithm would require the knowledge of the set of all minimal forbidden minors for the class of partial k-trees.

Acknowledgments. Derek G. Corneil wishes to express his appreciation of the hospitality of IMAG at the Université de Grenoble. Andrzej Proskurowski gratefully acknowledges the hospitality of the Royal Institute of Technology. The referees have provided helpful criticism, and in particular suggested an improved presentation of the results in § 3.

REFERENCES

[1] S. Arnborg, *Reduced state enumeration—Another algorithm for reliability evaluation*, IEEE Trans. Reliability, R-27 (1978), pp. 101–105.

[2] ———, *Efficient algorithms for combinatorial problems on graphs with bounded decomposability—A survey*, BIT, 25 (1985), pp. 2–33.

[3] S. Arnborg and A. Proskurowski, *Characterization and recognition of partial* 3-*trees*, this Journal, 7 (1986), pp. 305–314.

[4] ———, *Linear time algorithms for* NP-*hard problems on graphs embedded in k-trees*, TRITA-NA-8404, The Royal Institute of Technology, 1984.

[5] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, North-Holland, Amsterdam, 1976.

[6] C. J. Colbourn and A. Proskurowski, *Concurrent transmissions in broadcast networks*, Proc. Int. Conf. Automata, Languages, Programming, Springer-Verlag, Berlin–Heidelberg-New York, Lecture Notes in Computer Science, 172 (1984), pp. 128–136.

[7] D. G. Corneil and J. M. Keil, *A dynamic programming approach to the dominating set problem on k-trees*, this Journal, 8 (1987), to appear.

[8] A. M. Farley, *Networks immune to isolated failures*, Networks, 11 (1981), pp. 255–268.

[9] A. M. Farley and A. Proskurowski, *Networks immune to isolated line failures*, Networks, 12 (1982), pp. 393-403.

[10] M. R. Garey and D. S. Johnson, *Computers and Intractability*, W. H. Freeman, San Francisco, CA, 1979.

[11] P. C. Gilmore and A. J. Hoffman, *A characterization of comparability graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.

[12] E. M. Neufeldt and C. J. Colbourn, *The most reliable series-parallel networks*, Dept. of Computing Science, University of Saskatchewan, TR 83-7, 1983.

[13] A. Proskurowski, *Separating subgraphs in k-trees*: *Cables and caterpillars*, Discrete Math., 49 (1984), pp. 275–285.

[14] D. J. Rose, *Triangulated graphs and the elimination process*, J. Math. Anal. Appl., 32 (1970), pp. 597–609.

[15] ———, *On simple characterization of k-trees*, Discrete Math., 7 (1974), pp. 317–322.

[16] A. Wald and C. J. Colbourn, *Steiner trees, partial* 2-*trees, and minimum* IFI *networks*, Networks, 13 (1983), pp. 159–167.

[17] M. Yannakakis, *Computing the minimum fill-in is* NP-*complete*, this Journal, 2 (1981), pp. 77–79.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 2, April 1987

014

DISTRIBUTION OF THE MINIMUM CHANNEL WIDTH IN VLSI WIRING*

D. COPPERSMITH†, I. GOPAL† AND C. K. WONG†

Abstract. Suppose we have N terminals on one side of a wiring channel, and N on the other side, and we wish to achieve a given interconnection specified by a randomly chosen permutation function. We show that the minimum number of horizontal channels necessary is close to $N/2$ most of the time.

Key words. VLSI, wiring channel width, random walk

AMS(MOS) subject classifications. 60J15, 68A05

1. Introduction. The problem of channel wiring has become increasingly important in VLSI design. Specifically, the problem is one of interconnecting two sets of terminals, one set on each side of a wiring channel, and to accomplish this interconnection while optimizing some objective function. Typically this objective function is the channel *width* or the number of horizontal tracks necessary in the channel.

Several previous efforts have been directed towards obtaining minimum or near-minimum width solutions for some given problem instance [1], [2]. However, it is often of great use to the designer to obtain some idea of the minimum width without a complete specification of the problem. Specifically, estimates of the distribution of the minimum width can be of value in making layout and global routing decisions.

In this paper, we develop such estimates for a specific wiring model. We consider a two layer wiring channel in which vertical wire segments are restricted to one layer and horizontal wire segments to another, vertical and horizontal wire segments being joined by means of via holes. All connections involve exactly one terminal on each side of the wiring channel. Thus, if the terminals on the upper side of the channel are labeled (from left to right) $a_1, a_2, \cdots, a_N$, and the terminals on the lower side are labeled $b_1, b_2, \cdots, b_N$, the connections are completely specified by the permutation function $\pi : \{1, \cdots, N\} \rightarrow \{1, \cdots, N\}$ which specifies that a_i is connected to $b_{\pi(i)}$ for every i, $1 \leqq i \leqq N$.

Clearly, to complete the specification of a problem instance it is necessary to specify the embedding of terminals, i.e., the positioning of the terminals relative to each other. We consider, in this paper, two very specific types of embeddings. The first is the "uniform" embedding shown in Fig. 1 in which terminal b_1 is positioned between a_1 and a_2, b_2 is positioned between a_2 and a_3 and so on, with b_N lying to the right of a_N. The second type is the "optimal" embedding, where an "optimal" embedding for a given permutation function π is defined as the embedding which allows a wiring of globally minimum width, the global minimum now being taken over all possible wirings and over all possible embeddings. Figure 2 shows an example of an "optimal" embedding. A previous paper [3] presents an algorithm to find this optimal embedding for any specified permutation.

We note that, for both embeddings, there are no "vertical constraints" of the type discussed in [5] and thus the minimum channel width is simply the maximum crossing number [3]. (The crossing number at some vertical line through the wiring channel is the number of connections with the left terminal on or to the left of the vertical line and the right terminal to the right of the vertical line. The maximum crossing number is the maximum over all such vertical lines.)

* Received by the editors June 30, 1982; accepted for publication (in revised form) October 30, 1986.

† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

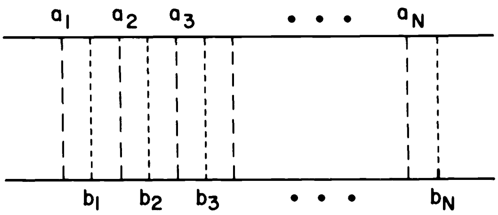


FIG. 1. *A uniform embedding.*

For both types of embeddings, we investigate the distribution of the minimum channel width over all possible permutation functions, each of the $N!$ permutations being assumed equally likely.

We shall show that, for most permutations, under either embedding, the minimum width is very close to $N/2$. In fact, for each ε there is a c such that with probability $1 - \varepsilon$, the width is between $(N/2) - c\sqrt{N}$ and $(N/2) + c\sqrt{N}$. The argument for the upper bound will be by comparison with a random walk with fixed end points while the lower bound will follow from a simple geometric argument. We also show that, for both embeddings, the expected value of the minimum width is at least $(N - 1)/2$.

2. Mathematical formulations.

DEFINITION. A *path* P of length L is a sequence of points $(0, P_0)$, $(1, P_1)$, $\cdots$, (L, P_L) in the plane integer lattice. We consider only finite paths here.

DEFINITION. A *generalized random walk* W is a collection of paths of length L, along with a probability density, such that

(1) Each path starts at $(0, y)$ for some *even* integer y;

(2) Each step in a path is from (x, y) to $(x + 1, y + 1)$ or to $(x + 1, y - 1)$ for some integers x, y;

(3) There is an initial density $i(y)$, $0 \leqq i(y) \leqq 1$, $\sum i(y) = 1$, and a set of transition probabilities $t(x, y) = t(x, y, W)$, $0 \leqq t(x, y) \leqq 1$, such that the probability associated with a path Y along the vertices $(0, Y_0)$, $(1, Y_1)$, $(2, Y_2)$, $\cdots$, (L, Y_L) is given by

$$\operatorname{prob}(Y) = i(Y_0) \prod_{0 \leqq x < L} u(x, Y_x, Y_{x+1}),$$

where

$$u(x, y, y+1) = t(x, y),$$
$$u(x, y, y-1) = 1 - t(x, y)$$
$$u(x, y, y') = 0 \quad \text{if } |y - y'| \neq 1.$$

In words, we start with some initial density $i(y) = i(y, W)$ of positions along the y-axis (even integers only). Then from any given point (x, y) we proceed to the right and up

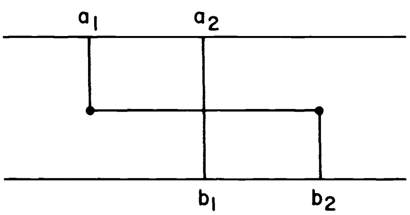


FIG. 2. *An optimal embedding.*

(to $(x+1, y+1)$) with probability $t(x, y)$, or to the right and down (to $(x+1, y-1)$) with probability $1 - t(x, y)$, independent of the history of our path to the left of point (x, y). We continue to position (L, y') for some y'.

The usual random walk starts with the initial density concentrated at 0 ($i(0) = 1$), and has uniform transition probabilities ($t(x, y) = \frac{1}{2}$, usually).

If we have two generalized random walks, W and W', both of length L, we say that *W majorizes W'* in the event that, given any path P of length L, $(0, P_0)$, $(1, P_1)$, $\cdots$, (L, P_L), the probability that a path Y in W lies below P (i.e., $Y_X \leqq P_x$, $x = 0, 1, \cdots, L$), is no greater than the probability that a path in W' lies below P.

3. Bounds on minimum channel width.

LEMMA 1. *Given two generalized random walks W and W' of length L, if the initial density $i(y, W)$ majorizes the initial density $i(y, W')$ (in the sense that $\Sigma_{y \leqq j} i(y, W) \leqq \Sigma_{y \leqq j} i(y, W')$ for all j), and if $t(x, y, W) \geqq t(x, y, W')$ for all (x, y), then W majorizes W'.*

Proof of Lemma 1. Let $g(x, y, W, P)$ denote the proportion of paths Y in W whose initial segments (up to and including x) lie below P, and with $Y_x \leqq y$. By hypothesis on $i(y, W)$ and $i(y, W')$, we have $g(0, y, W, P) \leqq g(0, y, W', P)$ for all y and P.

Now use induction on x. Suppose first $y \leqq P_x$, and that x and y have the same parity. Set

$$\begin{aligned} G &= g(x, y, W, P), & G' &= g(x, y, W', P), \\ A &= g(x-1, y+1, W, P), & A' &= g(x-1, y+1, W', P), \\ B &= g(x-1, y-1, W, P), & B' &= g(x-1, y-1, W', P), \\ T &= t(x-1, y+1, W), & T' &= t(x-1, y+1, W'). \end{aligned}$$

Then, from $T \geqq T'$, $A' \geqq A$, $A' \geqq B' \geqq B$, and the relations

$$G = B + (1-T)(A-B) = TB + (1-T)A,$$

$$G' = B' + (1-T')(A'-B') = T'B' + (1-T')A',$$

we calculate

$$G' - G = (1-T)(A'-A) + T(B'-B) + (T-T')(A'-B') \geqq 0.$$

If y and x are of different parity, then $g(x, y, W, P) = g(x, y-1, W, P)$. Also, if $y > P_x$, then $g(x, y, W, P) = g(x, P_x, W, P)$. This completes the induction. □

In particular, letting P range over the constant horizontal paths, we see that if W majorizes W', then the distribution of maximum heights of the paths in W majorizes that of W'. If W majorizes W', then the expected maximum height of a path in W, is at least as large as the expected maximum height of a path in W'.

We use this lemma to prove an upper bound on the distribution of widths for the uniform embedding.

THEOREM 1. *Let N be even. The width of a random permutation of N terminals, under a uniform embedding, is majorized by the maximum height of a random walk from $(N/2, N/2)$ to $(3N/2, N/2)$. Thus, for each ε, there is a c such that $(1-\varepsilon)$ of the paths have width less than $N/2 + c\sqrt{N}$.*

Proof. For $0 \leqq k \leqq 2N$, let $f(k)$ be the number of wires crossing a vertical line drawn between $a_{\lceil (k+1)/2 \rceil}$ and $b_{\lceil k/2 \rceil}$. The width of the channel will be $\max_k f(k)$ as this is exactly the maximum crossing number. It turns out that this function $f(k)$ leads naturally to a generalized random walk W'. The hardest thing to check is that the probability of going up from $(k, f(k))$ to $(k+1, f(k)+1)$, i.e., $\text{Prob}\{f(k+1) = f(k)+1\}$, depends only on

k and $f(k)$, not on the previous history of f. But this is assured by the nature of random permutations, and in fact we have that

$$\begin{aligned} t(x, y, W') &= (2N-x-y)/(2N-x) \quad \text{if } x \text{ is even,} \\ &= (2N-x-y)/(2N-x+1) \quad \text{if } x \text{ is odd.} \end{aligned}$$

We will consider only that portion of the path W' lying between $x = N/2$ and $x = 3N/2$. We will consider also the generalized random walk W, which assigns uniform probability to all paths running from $(N/2, N/2)$ to $(3N/2, N/2)$. We will show that W majorizes W'. Then, for the portions of W' where $0 \leqq x < N/2$ or $3N/2 < x \leqq 2N$, it is immediate that $f(x) < N/2$. Thus it is only the middle portion, $N/2 \leqq x \leqq 3N/2$, which is of interest.

It turns out to be easier to start from the middle and work outwards in both directions. Consider first the density of y-coordinates at the point $x = N$ for both W and W'. (We call this density $i(y, W)$, and let our generalized random walk start at N instead of at 0.) First,

$$i(y, W) = \binom{N/2}{y/2}^2 \Big/ \binom{N}{N/2},$$

since in order to go from $(N/2, N/2)$ to (N, y), we must have taken $y/2$ steps up and $(N/2) - y/2$ steps down, so that there are $\binom{N/2}{y/2}$ paths from $(N/2, N/2)$ to (N, y); there are $\binom{N/2}{y/2}$ paths from (N, y) to $(3N/2, N/2)$; and there are $\binom{N}{N/2}$ paths from $(N/2, N/2)$ to $(3N/2, N/2)$.

Consider $i(y, W')$, on the other hand. Of the $N!$ permutations, there are $\binom{N/2}{y/2}$ ways of choosing $y/2$ nodes from among the top left nodes $a_1, \cdots, a_{N/2}$ to connect to nodes on the bottom right; $\binom{N/2}{y/2}$ ways of choosing $y/2$ nodes from among the top right to connect to nodes on the bottom left; $(N/2)!$ ways of arranging the $N/2$ wires now leading to the bottom left; and $(N/2)!$ ways of arranging the $N/2$ wires leading to the bottom right. Thus

$$i(y, W') = \binom{N/2}{y/2}^2 [(N/2)!]^2/N! = i(y, W).$$

Since the initial densities are the same, each majorizes the other.

Let us proceed to the right, from $x = N$ to $x = 3N/2$. For the walk W, we find that

$$t(x, y, W) = (2N-x-y)/(3N-2x),$$

since of the $(3N/2 - x)$ steps remaining to reach $(3N/2, N/2)$, $(2N - x - y)/2$ must be upwards and $(N - x + y)/2$ must be downwards, and the uniform nature of W demands that we respect this ratio in deciding our next step.

On the other hand, for W' we have:

$$\begin{aligned} t(x, y, W') &= (2N-x-y)/(2N-x) \quad \text{if } x \text{ is even,} \\ &= (2N-x-y)/(2N-x+1) \quad \text{if } x \text{ is odd.} \end{aligned}$$

This is seen (for x even) by saying that of the $(2N - x)/2$ nodes on the top right (a_i for $i > x/2$), there are $(2N - x - y)/2$ nodes which connect to nodes on the bottom right, and if $a_{x/2+1}$ is one of those, we will have $f(x + 1) = f(x) + 1$. Similar arguments hold for x odd.

Since $x \geqq N$, we have $2N - x + 1 > 2N - x \geqq 3N - 2x$, so that

$$t(x, y, W') \geqq t(x, y, W).$$

Thus, by our lemma, we have shown that the portion of W' with $N \leqq x \leqq 3N/2$ majorizes the corresponding portion of W.

We can view the portions of W and W' with $N \geqq x \geqq N/2$ as generalized random walks running backwards, and remark that the portions with $N \geqq x \geqq N/2$ and the portions with $N \leqq x \leqq 3N/2$ are independent, except for the value at $x = N$. For example, for W, any path from $(N/2, N/2)$ to (N, y) can be joined with any path from (N, y) to $(3N/2, N/2)$, and the probability of the entire path is just $i(y, W)$ times Prob{right-hand path $|(N, y)$} times Prob{left-hand path $|(N, y)$}.

Thus, the generalized random walk W majorizes the walk W'. In particular, the maximum height achieved by W majorizes that achieved by W'. But W is well understood. For example, we can apply the reflection principle [4] to show that the probability that a path in W reaches height $(N/2) + h$ is exactly

$$\binom{N}{(N/2)+h} \Big/ \binom{N}{N/2},$$

and this decays as $e^{-2h^2/N}$, so that for each ε we can find a c so that the probability of exceeding $(N/2) + c\sqrt{N}$ is less than ε. By the fact that W majorizes W', we have immediately the same statement about W', which is what we wanted. □

Remark. Clearly, the results of Theorem 1 apply to the optimal embedding. The next theorem provides a lower bound on the average width of the optimal embedding (and thus also on the uniform embedding).

THEOREM 2. *Under the optimal embedding, the average minimum width is at least* $(N-1)/2$.

Proof. Let π' be the "reverse" permutation of π, namely $\pi'(i) = N + 1 - \pi(i)$. (Note that this is not the "inverse.") This corresponds to flipping the bottom half of the wiring channel. A simple geometric argument shows that width (π) + width $(\pi') \geqq N - 1$. This will show a lower bound on the average is at least $(N-1)/2$.

Let there be given the optimal embeddings of π and π'. For simplicity, assume that terminal a_i lies directly above terminal b_j only when they are connected by a wire $(j = \pi(i))$. (We can always do this without increasing the width of the embedding.) By continuity, we argue that there is a vertical line x on the wiring channel realizing π, and a vertical line x' on the wiring channel realizing π', such that:

(a) If x lies between a_i and a_{i+1} on the top of π, then x' lies between a'_i and a'_{i+1} on the top of π';

(b) If x lies between b_j and b_{j+1} on the bottom of π, then x' lies between b'_{N+1-j} and b'_{N-j} on the bottom of π';

(c) If x lies exactly on terminals a_i and b_j on π, then x' lies exactly on terminals a'_i and b'_{N+1-j} on π'.

Then an easy argument shows that the number of wires crossing x, plus the number of wires crossing x', is exactly N, unless x and x' both lie directly on wires (from a_i to b_j and from a'_i to b'_{N+1-j}, respectively), in which case it is exactly $N - 1$. In fact, the wire from a_k to $b_{\pi(k)}$ crosses x if and only if the wire from a'_k to $b'_{\pi'(k)}$ does not cross x', with the sole exception being $k = i$ and $j = \pi(k)$.

The width of each embedding is at least as big as its width at this point x or x'. So the widths of the two embeddings add to at least $N - 1$. In particular this is true of minimal embeddings. Thus width (π) + width $(\pi)' \geqq N - 1$. By summing over all π, and dividing by two we get our desired result. □

Remark. Combining this proof with the previous result on upper bounds on the distribution of widths, we get a lower bound on the distribution of widths under either

TABLE 1

Minimum width	Fraction of experimental trials
0	0
1	0
2	0
3	0
4	0
5	0.04
6	0.16
7	0.25
8	0.3
9	0.21
10	0.04
11	0
12	0
13	0
14	0
15	0

type of embedding. That is, the probability that a width is less than $(N/2) - 1 - c\sqrt{N}$, is no more than the probability that a width (of the reverse permutation) is greater than $N/2 + c\sqrt{N}$, since the two must sum to at least $N - 1$.

4. Experimental results. In a previous paper [3], we presented an algorithm to find an optimal embedding and the corresponding minimum width wiring for a given permutation function. To verify our theoretical results of § 3 we implemented the algorithm in APL and ran it on 100 randomly generated permutation functions on 15 terminals. Table 1 reports the distribution of the minimum width obtained.

As can be seen the distribution is sharply peaked, with a mean of around 8, which would tend to corroborate our theoretical results.

REFERENCES

[1] A. HASHIMOTO AND J. STEVENS, *Wire routing by optimizing channel assignment within large apertures*, Proc. 8th Design Automation Workshop, June 1971, pp. 115–169.

[2] D. N. DEUTSCH, *A "dogleg" channel router*, Proc. 13th Design Automation Workshop, 1976, pp. 425–433.

[3] I. GOPAL, D. COPPERSMITH AND C. K. WONG, *Optimal wiring of movable terminals*, IEEE Trans. Comp., C-32, (1983), pp. 845–858.

[4] W. FELLER, *An Introduction to Probability Theory and Its Applications*, John Wiley, New York, 1968.

[5] T. KAWAMOTO AND Y. KAJITANI, *The minimum width routing of a 2-row 2-layer polycell-layout*, Proc. 16th Design Automation Conference, 1979, pp. 290–296.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

TENSOR EQUIVALENTS FOR SOLUTION OF LINEAR SYSTEMS: A PARALLEL ALGORITHM*

JOHN DE PILLIS†

Abstract. In this paper, we develop a stationary iterative method to find the solution vector x for the invertible $n \times n$ linear system

$$Ax=(I-B)x=f. \tag{1.1}$$

(I and I_k represent the appropriate $k \times k$ identity matrix.) We find x by replacing (1.1) with the equivalent system

$$\tilde{A}\tilde{x}=(I-\tilde{B})\tilde{x}=\tilde{f}, \qquad \tilde{x}=x_0 \otimes x. \tag{1.2}$$

Solution vector x of (1.1) will be "easy to extract" from the solution of (1.2) since $\tilde{x} = x_0 \otimes x$ is always a decomposable tensor. For *any* quadratic polynomial where $\varphi(1) = 1$, we may construct tensor iteration matrix $\tilde{B}$ of (1.2) whose eigenvalues $\pm\lambda$, are all determined by φ according to the equation $(*)\lambda^2 = \varphi(\mu)$ where μ runs over the eigenvalues of B in (1.1). With the ability to shape the spectrum of $\tilde{B}$ as per $(*)$, we develop an optimal stationary iterative algorithm to solve (1.2) in the special case when the spectrum of A in (1.1) is real. The algorithm is further enhanced if its parallelism is exploited.

Key words. parallel algorithm, iteration, sparse matrix, tensor, linear system

AMS(MOS) subject classifications. 65F10, 68B99, 15A69

1. Introduction.

1.1 Equivalent systems. Systems (1.1) and (1.2) have solutions, both of which depend on the same vector x. We call these systems *equivalent* whenever

$$\tilde{x}=(x_0 \otimes x), \qquad \tilde{B}=P\otimes I+Q\otimes B \quad \text{and} \quad \tilde{f}=f_0 \otimes f \tag{1.3}$$

where 2×2 matrices Q and P and 2-vectors x_0 and f_0 are to be determined in advance. (*Note*: Tensor $\otimes$ definitions and basic properties are provided in the following sections.) Substitution of (1.3) into (1.2) produces the equivalent tensor system

$$(I_{2n}-\tilde{B})(x_0 \otimes x)=f_0 \otimes f \quad \text{where } \tilde{B}=P\otimes I+Q\otimes B. \tag{1.4}$$

Having found tensor vector $(x_0 \otimes x)$ in (1.4), we will "factor out" vector x_0 to finally obtain solution vector x of (1.1).

Remark (preserving decomposable tensors). From (1.4), we see that tensor iteration matrix $\tilde{B}$ must preserve decomposable tensors, i.e., $I_{2n} - \tilde{B}$ sends decomposable $(x_0 \otimes x)$ to decomposable $(f_0 \otimes f)$. Tensor definitions follow.

Why is system (1.4) (a special case of (1.2)) easier to work with than (1.1)? As we shall see in (4.3), we may link the spectra of $\tilde{B}$ and B by means of an arbitrary polynomial φ of degree two—the only constraint on φ is that $\varphi(1) = 1$.

Here is what we mean by the "linking" of $\sigma(B)$ with $\sigma(\tilde{B})$: We are given linear system (1.1) and suppose φ is *any* quadratic polynomial such that $\varphi(1) = 1$. Then 2×2 matrices Q and P may be chosen so that $\tilde{B}$ of (1.3) has spectrum

$$\sigma(\tilde{B})^2=\varphi(\sigma(B)); \tag{1.5a}$$

* Received by the editors February 10, 1986; accepted for publication (in revised form) May 8, 1986.

† Department of Mathematics and Computer Science, University of California, Riverside, California 92521.
This work was partially supported by the Air Force Office of Scientific Research under grant AFOSR-84-0380 and the IBM Scientific Center, Bergen, Norway.

that is, each $\mu \in \sigma(B)$ defines the pair of values $\pm\lambda \in \sigma(\tilde{B})$ as follows:

$$\lambda^2 = \varphi(\mu), \quad \text{where } \varphi(1) = 1. \tag{1.5b}$$

Remark. The linking conditions (1.5a, b), are reminiscent of the well-known eigenvalue equation (1.7) of the SOR method for consistently ordered systems. Here is a brief description (cf. [22], [23]). In the SOR method, linear systems involving invertible A of (1.1) are put into the form

$$Ax = (I_n - L - U)x = f. \tag{1.6a}$$

A is said to be *consistently ordered* relative to the splitting (1.6a) whenever the following spectral invariance holds: for all scalars α, $\beta \neq 0$,

$$\sigma(\alpha L + \alpha^{-1}U) = \sigma(\beta L + \beta^{-1}U). \tag{1.6b}$$

If we set $B = (L + U)$ in (1.1), we obtain (1.6a). In the terminology of (1.3), the SOR method induces a linear system which is equivalent to (1.1). To see this, substitute into (1.2) as follows:

$$\tilde{x} = x,$$
$$\tilde{B} = \tilde{B}_\omega = (I_n - \omega L)^{-1}[(1-\omega)I_n + \omega U], \qquad 0 < \omega < 2,$$
$$\tilde{f} = (I_n - \omega L)^{-1}f.$$

A principal consequence of consistent ordering is that the spectra of the two iteration matrices, B and $\tilde{B}$, are linked in a manner similar to (1.5). In fact, we have that $\mu \in \sigma(B)$ and $\lambda \in \sigma(\tilde{B}_\omega)$ are linked by the well-known functional relation

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2. \tag{1.7}$$

Equation (1.7) behaves like (1.5b) in that $\mu = 1$ implies $\lambda = 1$.

In (1.5), there is a good deal of freedom in "molding" $\sigma(\tilde{B})$ once we know the geometry of $\sigma(B)$. This paper considers one special case: when $\sigma(A)$, the eigenvalues of A, are real and straddle the origin (equivalently, eigenvalues $\sigma(B) = \sigma(I - A)$ are real and straddle the point $z = 1$.) More exactly, we shall assume our matrices A have the property

$$\sigma(A) = \sigma(I - B) \subset [a, b] \cup [c, d], \qquad a \leqq b < 0 < c \leqq d. \tag{1.8}$$

Systems (1.1) with condition (1.8) include all symmetric Hermitian matrices A. This is discussed in § 8 along with a brief survey of the literature.

The success of stationary iterative methods depends entirely on the spectrum of an iteration matrix like B of (1.1) or $\tilde{B}$ of (1.4). We briefly describe the role of iteration matrices in the creation of solution sequences and in measuring convergence rates of these sequences.

1.2. Convergence rates of stationary methods. Consider the typical invertible $n \times n$ linear system, like that of (1.1) or (1.4), which is of the form $(I - G)y = g$. Then arbitrary initial n-vector y_0 produces the sequence

$$y_k = Gy_{k-1} + g, \qquad k = 1, 2, 3, \cdots. \tag{1.9}$$

Sequence y_k of (1.9) converges to solution vector y where $(I - G)y = g$, if and only if $\rho(G)$, the spectral radius of iteration matrix G, is less than one. The speed of convergence of (1.9) increases as ρ decreases: The *asymptotic convergence rate* R_y, of the sequence (1.9), is quantitized once we define

$$R_y = -\log(\rho(G)). \tag{1.10}$$

An interpretation of the convergence rate (1.10) is this: $(1/R_y)$ is, asymptotically, the number of iterations which suffice to produce an added (decimal) place of accuracy in y_k; cf., Varga [23]. Denote this number by *step_count* so that

$$\text{(1.11)} \qquad step\_count = \frac{-1}{\log 10(\rho(G))}.$$

It follows from (1.11) that fewer iterates will suffice to produce added accuracy as $\rho(G)$ becomes smaller.

1.3. Parallel processing. The question is not yet settled as to what the "optimal" parallel architecture should be. For example, should our parallel machine allow message passing as is the case in a hypercube topology? Or should all processors share a common memory? In the former case, there are time delays in bundling the messages in order to ship them from processor to processor; in the latter case, badly timed overwriting of global variables (memory contention) is a potential problem.

In developing algorithms for parallel machines, we often look for parallelism in current serial algorithms. One example is the computation of the inner product which involves simultaneous multiplication of scalar pairs. For another example, note that once the scalar multipliers are known in Gaussian elimination, the sequence of row operations which produce the zeros below the diagonal may just as well proceed simultaneously, or in parallel.

Our algorithm involves several matrix multiplications and vector linear combinations. (See the pseudocode in § 6.) Therefore we may look for parallelism within these standard matrix manipulations.

But there is a true parallelism in our algorithm that is indicated by (6.15) and (6.16) of the pseudocode. The computation of co-sequences $\{v_k\}$ and $\{w_k\}$ may proceed independently within the same time frame. On a serial machine, of course, (6.15) and (6.16) must be computed in sequence. Moreover, our theory has it that p separate co-sequences result if the matrices P and Q are $p \times p$. In this work, we consider only the special case $p = 2$.

1.4. Structure of this paper.

- Section 2 provides basic information and definitions on tensor products of matrices.
- Section 3 characterizes the spectrum of matrices that are sums of tensors. In particular, vectors x_0 and f_0 are described in (3.11) for any given system (1.1). The matrix $\tilde{B}$ of (1.2) is characterized in (3.2).
- Section 4 is devoted to characterizing the spectrum of $\tilde{B}$ of (1.2). In fact, we see in Theorem 4.1 that any quadratic polynomial $\varphi(1) = 1$ defines matrices Q and P in (1.2) which, in turn, leads to the linking condition (1.5).
- Section 5 proves the optimality of the our algorithm, which produces an equivalent tensor system (1.4) and follows it with a two-part acceleration; cf. [4], [14]. In our case, optimality is with respect to the two-part splitting acceleration. That is to say, the *final* spectral radius $\rho(G)$ of (1.10) is determined by the zero-centered ellipse which contains all the eigenvalues of iteration matrix $\tilde{B}$ given by (1.4). As required in the theory of two-part splitting, we construct the enveloping ellipse for $\sigma(\tilde{B})$ and denote its four vertices in the complex plane by

$$\text{(1.12a)} \qquad iV, -iV, H, -H \quad \text{where } H, V \geqq 0.$$

Then the sequence

$$\text{(1.12b)} \qquad \tilde{y}_k \rightarrow (x_0 \otimes x)$$

converges to solution vector $(x_0 \otimes x)$ of (1.4), with accelerated convergence rate

$$\text{(1.12c)} \qquad R_{\tilde{y}} = -\log(\rho(H, V))$$

where

$$\rho(H, V)=\frac{\sqrt{1+V^2-H^2}-1}{V-H}. \tag{1.12d}$$

Different semi-axes H, V, of course, produce different values of $\rho(H, V)$ in (1.12d) which, in turn, produce different convergence rates $R_{\tilde{y}}$ (1.12c). The optimal, i.e., smallest, $\rho = \rho(H, V)$ of (1.12d) will be shown to occur when $H = 0$. Thus, among the family of tensor iteration matrices $\tilde{B}$ in (1.2), the optimal one with respect to two-part acceleration turns out to be exactly the $\tilde{B}$ with pure imaginary spectrum. (See Lemma 5.5.)

- Section 6 presents a pseudo code for the implementation of the algorithm when matrix A of (1.1) has straight-line spectrum (1.8). The parallel nature of our algorithm can be seen, for example, in (6.15) and (6.16) which entail two simultaneous matrix-vector products. Moreover, the stopping criteria, (6.17a) and (6.17b), require another pair of simultaneous or parallel matrix-vector products. Also, the sequences $\{v_k\}$ and $\{w_k\}$, account for several vector linear combinations which can be executed in parallel.
- Section 7 offers concluding remarks on the algorithm including relation of our techniques to condition number.
- Section 8 presents a brief overview of the literature based on methods which allow or do not allow $\sigma(A)$ to straddle the origin.

2. Tensor fundamentals. We recall some definitions and properties of tensor (Kronecker) products of vectors.

DEFINITION 2.1. *$X \otimes Y$, the decomposable tensor product of X with Y*: Given $p \times q$ matrix X and $r \times s$ matrix Y, where $X = (a_{i,j})$, $i = 1, 2, \cdots p, j = 1, 2, \cdots, q$. Then $X \otimes Y$ is the $pr \times qs$ matrix defined by

$$X \otimes Y = (a_{i,j}Y), \qquad i = 1, 2, \cdots p, \quad j = 1, 2, \cdots, q. \tag{2.1}$$

If we use the matrix product in our definition of tensor product, we obtain the equivalent definition.

DEFINITION 2.1′. *$X \otimes Y$, the decomposable tensor product of X with Y*: Given $p \times q$ matrix X and $r \times s$ matrix Y, then $X \otimes Y$, is the linear transformation defined on all $q \times s$ matrices Z as follows:

$$X \otimes Y : Z \rightarrow XZY^t \tag{2.1′}$$

where Y^t is the transpose matrix of Y.

If $q \times s$ matrix Z is identified with the $qs \times 1$ column matrix (write the entries of Z in lexicographic order), then it is fairly direct to show the equivalence of (2.1) and (2.1′).

DEFINITION 2.2. *$V \otimes W$, the tensor product of vector spaces V and W*, is the linear span of all the individual decomposable tensors $X \otimes Y$ where $X \in V$ and $Y \in W$.

Property 2.3. Let matrices X, X', Y, Y' be compatibly dimensioned in the sense that matrix products (XX') and (YY') exist. Then the product of decomposable tensors yields another decomposable tensor, viz.,

$$(X \otimes Y)(X' \otimes Y') = XX' \otimes YY'. \tag{2.2}$$

In the special case that matrices X and Y are square, $\sigma(X)$ and $\sigma(Y)$, the spectra of X and Y, respectively, are easily related to $\sigma(X \otimes Y)$. In fact, $\sigma(X \otimes Y) = \sigma(X) \cdot \sigma(Y)$. More precisely, we have

Property 2.4. Suppose $m \times m$ matrix X and $p \times p$ matrix Y have respective (distinct) eigenvalues

$$\sigma(X) = \{\mu_1, \mu_2, \cdots, \mu_m'\}, \qquad m' \leqq m$$

and

$$\sigma(Y)=\{\lambda_1,\lambda_2,\cdots,\lambda_{p'}'\},\qquad p'\leqq p.$$

Then $X\otimes Y$ has spectrum

$$\sigma(X\otimes Y)=\{(\mu_i\cdot\lambda_j)\},\qquad i=1,2,\cdots,m',\quad j=1,2,\cdots,p'. \tag{2.3}$$

A consequence of commutativity of the eigenvalues in (2.3), is that for any matrices X and Y,

$$\sigma(X\otimes Y)=\sigma(Y\otimes X). \tag{2.4a}$$

The form of (2.3) suggests the notation $\sigma(X)\cdot\sigma(Y)$ for $\sigma(X\otimes Y)$. Accordingly, (2.3) and (2.4a) together can be rewritten

$$\sigma(X\otimes Y)=\sigma(Y\otimes X)=\sigma(X)\cdot\sigma(Y). \tag{2.4b}$$

The identity (2.4a) extends from single decomposable tensors $X\otimes Y$ to any linear combination of decomposable tensors. That is,

$$\sigma\left(\sum_i X_i\otimes Y_i\right)=\sigma\left(\sum_i Y_i\otimes X_i\right). \tag{2.5}$$

This concludes our brief survey of tensor properties of matrices.

3. The spectrum of tensor sums. We now develop an eigenvalue characterization for sums of decomposable tensors $L_i\otimes X_i$ when the X_i's can simultaneously be put into upper triangular form. Here is that theorem now.

THEOREM 3.1. *Given arbitrary $p\times p$ matrices $\{L_i\}$, $i=0,1,2,\cdots,k$. Given the $n\times n$ matrices $\{X_i\}$, $i=0,1,2,\cdots,k$ and fixed invertible $n\times n$ matrix S which transforms each X_i to upper triangular form. That is,*

$$SX_iS^{-1}=\begin{bmatrix}\mu_{i,1} & * & * & \cdots & * \\ 0 & \mu_{i,2} & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & * \\ 0 & 0 & 0 & 0 & \mu_{i,n}\end{bmatrix},\qquad i=0,1,2,\cdots,k. \tag{3.1}$$

Then the spectrum of the tensor sum

$$\sigma\left(\sum_{i=0}^{k} L_i\otimes X_i\right) \tag{3.2}$$

is the union of the spectra of certain $p\times p$ matrices, viz.,

$$\bigcup_{j=1}^{n}\sigma\left(\sum_{i=0}^{k}\mu_{i,j}L_i\right). \tag{3.3}$$

Proof. We pass from (3.2) to (3.3) as follows:

$$\sigma\left(\sum_{i=0}^{k} L_i\otimes X_i\right)=\sigma\left(\sum_{i=0}^{k} X_i\otimes L_i\right)\quad\text{from (2.5)}$$

$$=\sigma\left((S\otimes I)\left(\sum_{i=0}^{k} X_i\otimes L_i\right)(S^{-1}\otimes I)\right)$$

$$=\sigma\left(\sum_{i=0}^{k} SX_iS^{-1}\otimes L_i\right)\quad\text{from (2.2).} \tag{3.4}$$

Now from (2.1) and (3.1), we recognize the matrix in (3.4) to be the $n \times n$ upper triangular block matrix

$$(3.5) \qquad \begin{bmatrix} \sum\limits_{i=0}^{k} \mu_{i,1} L_i & * & * & \cdots & * \\ 0 & \sum\limits_{i=0}^{k} \mu_{i,2} L_i & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & * \\ 0 & 0 & 0 & 0 & \sum\limits_{i=0}^{k} \mu_{i,n} L_i \end{bmatrix},$$

where each block is $p \times p$. The spectrum of this matrix is the union of the eigenvalues of the diagonal blocks; that is, the eigenvalues of (3.5) are given by

$$\bigcup_{j=1}^{n} \sigma\left(\sum_{i=0}^{k} \mu_{i,j} L_i \right).$$

The chain of equalities leading to (3.4) shows that (3.2) and (3.3) are equivalent. The theorem is proved. □

Theorem 3.1 assumes the following form in an important special case when $k = 1$.

COROLLARY 3.2. *Given $n \times n$ matrices I and B where*

$$\sigma(B) = \{\mu_1, \mu_2, \cdots, \mu_n\},$$

for any $p \times p$ matrices P, Q, the tensor sum $(P \otimes I) + (Q \otimes B)$ has eigenvalues

$$(3.6) \qquad \sigma((P \otimes I) + (Q \otimes B)),$$

which is equal to the following union of eigenvalues:

$$(3.7) \qquad \bigcup_{j=1}^{n} \sigma(P + \mu_j Q).$$

Proof. Let $n \times n$ matrix S define the similarity transformation which brings matrix $B = I - A$ to Jordan normal form. That is,

$$S(B)^i S^{-1} = \begin{bmatrix} (\mu_1)^i & * & 0 & \cdots & 0 \\ 0 & (\mu_2)^i & * & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & * \\ 0 & 0 & 0 & 0 & (\mu_n)^i \end{bmatrix}, \qquad i = 0, 1,$$

which assures us that hypothesis (3.1) is satisfied; Theorem 3.1 may now be invoked. Accordingly, displays (3.6) and (3.7) are just restatements and special cases of (3.2) and (3.3), respectively. This proves the corollary. □

Remark (Embedding vector x in a decomposable vector). What form may tensor iteration matrix $\tilde{B}$ take so that for all $x \in V_n$, we have $(I_{np} - \tilde{B})(x_0 \otimes x) = (f_0 \otimes Ax)$. That is, when does $I_{np} - \tilde{B}$ send n-dimensional (decomposable) subspace $x_0 \otimes V_n$ to (decomposable) subspace $f_0 \otimes AV_n$. This condition is formulated by (3.11); the details follow in the next theorem.

THEOREM 3.3. *Given invertible $n \times n$ matrix $A = (I - B)$ and given any $p \times p$ matrices P, Q where*

$$(3.8a) \qquad 1 \in \sigma(P + Q);$$

that is, for some $x_0 \in V_p$ *we have*

$$(P+Q)x_0 = x_0, \tag{3.8b}$$

if we define $\tilde{B}$ *and* $f_0 \in V_2$ *by*

$$\tilde{B} = P \otimes I_n + Q \otimes B, \tag{3.9}$$

$$f_0 = Qx_0, \tag{3.10}$$

then for all $x \in V_n$, *we have the tensor equation*

$$(I_{pn} - \tilde{B})(x_0 \otimes x) = (f_0 \otimes Ax). \tag{3.11}$$

Proof. We establish the validity of (3.11) with the following equalities:

$$(I_{pn} - \tilde{B})(x_0 \otimes x) = (x_0 \otimes x) - \tilde{B}(x_0 \otimes x)$$

which, from (3.9),

$$\begin{aligned} &= (x_0 \otimes x) - (P \otimes I_n + Q \otimes B)(x_0 \otimes x) \\ &= (I_p - P)x_0 \otimes x - Qx_0 \otimes Bx \\ &= (I_p - P)x_0 \otimes x - Qx_0 \otimes (x - Ax) \\ &= (I_p - P - Q)x_0 \otimes x + Qx_0 \otimes Ax \\ &= Qx_0 \otimes Ax \quad \text{from (3.8b)} \\ &= f_0 \otimes Ax \quad \text{from (3.10).} \end{aligned}$$

This ends the proof. □

Remark. The theory of k-summability developed by W. Niethammer and R. Varga starts with systems of the form (1.1). Then iteration matrix B induces the k-part (acceleration) sequence

$$y_m = (\mu_0 B + \mu_1)y_{m-1} + \mu_2 y_{m-2} + \cdots + \mu_k y_{m-k} + \mu_0 f, \tag{3.12}$$

for arbitrary initial vectors $y_0, \cdots, y_{k-1}$ where $m = k+1, k+2, \cdots$, cf. [6], [7], [15]. Now the sequence (3.12) can be viewed as a special case of (2.9). In fact, write

$$\begin{bmatrix} y_m \\ y_{m-1} \\ \vdots \\ y_{m-k} \end{bmatrix} = \begin{bmatrix} \mu_0 B + \mu_1 & \mu_2 I & * & \cdots & \mu_{m-k} I \\ I & 0 & 0 & \cdots & 0 \\ 0 & I & * & \cdots & * \\ \vdots & \vdots & \vdots & 0 & * \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & I & 0 \end{bmatrix} \cdot \begin{bmatrix} y_{m-1} \\ y_{m-2} \\ \vdots \\ y_{m-k-1} \end{bmatrix} + \begin{bmatrix} \mu_0 f \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.13}$$

Now observe that matrix multiplication in the top row of (3.13) results in (3.12). At the same time, (3.13) is exactly of the form (1.9) with iteration matrix $G = P \otimes I + Q \otimes B$ where, from (2.1), we verify that

$$P = \begin{bmatrix} \mu_1 & \mu_2 & * & \cdots & \mu_{m-k} \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & * & \cdots & * \\ \vdots & \vdots & \vdots & 0 & * \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} \mu_0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & 0 & * \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We see that Q, unlike P, has fixed rank; in the theory of summability, the rank of $k \times k$

matrix Q is always one. This accounts for the fact that in the k-part sequence (3.12), the iteration matrix B appears only once while scalars μ_k are attached to the coefficients of all other iterate vectors.

4. The spectrum of $\tilde{B} = (P \otimes I + Q \otimes B)$. In this section, we show that for any $\lambda \in \sigma(\tilde{B})$, where $\tilde{B} = (P \otimes I + Q \otimes B)$ is given by (3.9), there exists $\mu \in \sigma(B)$ such that $\lambda^2 = \varphi(\mu)$, where φ is a quadratic polynomial that is *arbitrary* except for the fact that $\varphi(1) = 1$, which is a consequence of (3.8a) or (3.8b).

Here is the theorem that produces the 2×2 matrices Q and P to match the pregiven quadratic φ.

THEOREM 4.1. *Given the quadratic polynomial with arbitrary roots r, $s \neq 1$ where $\varphi(1) = 1$, i.e.,*

$$\varphi(z) = \frac{(z-r)(z-s)}{(1-r)(1-s)}, \tag{4.1}$$

then 2×2 matrices Q, P, may be chosen so that φ provides a correspondence between scalars λ and μ, where

$$\lambda \in \sigma(\tilde{B}), \qquad \tilde{B} = (P \otimes I + Q \otimes B) \tag{4.2a}$$

and

$$\mu \in \sigma(B). \tag{4.2b}$$

The precise relationship between λ and μ is given by the equation

$$\lambda^2 = \varphi(\mu) \tag{4.3a}$$

(*equivalently*)

$$\sigma(\tilde{B})^2 = \varphi(\sigma(B)), \tag{4.3b}$$

whenever the 2×2 matrices

$$Q = \begin{bmatrix} q & q_{1,2} \\ q_{2,1} & -q \end{bmatrix} \quad \textit{and} \quad P = \begin{bmatrix} p & 0 \\ 0 & -p \end{bmatrix} \tag{4.4}$$

have their entries p, q, q_{12} and q_{21} defined in terms of the roots r and s as follows:

$$p = \left[\frac{rs}{(1-r)(1-s)}\right]^{1/2}, \tag{4.5a}$$

$$q = -\frac{p(r+s)}{2rs}, \quad \textit{and} \tag{4.5b}$$

$$q_{12}q_{21} = 1-(q+p)^2. \tag{4.5c}$$

Proof. The operator $\tilde{B} = (P \otimes I + Q \otimes B)$ given in (4.2a) is a special case of (3.6) once we replace L_0 with P and replace L_1 with Q. It follows from (3.7), therefore, that

$$\sigma(\tilde{B}) = \cup\, \sigma(\mu Q + P) : \mu \in \sigma(B).$$

Now $\lambda \in \sigma(\tilde{B}) = \sigma(\mu Q + P)$ if and only if

$$0 = \det(\mu Q + P - \lambda I) = \lambda^2 - (\mu q + p)^2 - \mu^2 q_{12} q_{21}. \tag{4.6}$$

To accommodate condition (3.8a), viz., $1 \in \sigma(Q + P)$, we see from direct computation on 2×2 matrices that $1 \in \sigma(Q + P)$ if and only if $0 = \det(Q + P - I)$ if and only if

$$q_{12}q_{21} = 1-(q+p)^2. \tag{4.7}$$

This proves (4.5c). Now use (4.5c) to replace the product $q_{12}q_{21}$ in (4.6) above. We then obtain

$$\lambda^2=(1-2qp-p^2)\mu^2+(2qp)\mu+p^2. \tag{4.8}$$

Our proof is done once we show that the μ-polynomial of (4.8) has the same *roots* as φ of hypothesis (4.1). Now the roots of (4.8) are not altered if we divide the RHS through by $(1-2qp-p^2)$ to produce the *monic* polynomial

$$\mu^2+\frac{2qp}{(1-2qp-p^2)}\cdot\mu+\frac{p^2}{(1-2qp-p^2)}=\mu^2-(r+s)\mu+(r\cdot s).$$

This last equality follows since, for monic polynomials, the negative sum of roots always forms the μ coefficient and the product of roots always produces the constant term. Equating coefficients of the last two μ-polynomials above, produces the equalities (4.5a) and (4.5b). In fact, the matching of coefficients tells us that

$$(1-2qp-p^2)=\frac{-2qp}{r+s}=\frac{p^2}{r\cdot s}.$$

The right-hand equality above implies that (4.5b) holds. Substitute the value of q from (4.5b) into the left-hand term above, equate to the third term, solve for p and (4.5a) finally results. Recall that condition (4.5c) was established with (4.7). With (4.5a, b, c) thus established, the proof of the theorem is complete. □

5. Optimality of the algorithm. Let us briefly review our progress:

- We have shown in (3.11) of Theorem 3.3 that our original linear system with solution vector x and iteration matrix B, can be replaced by an equivalent tensor system with solution vector $(x_0\otimes x)$. That is, given $n\times n$ matrix $A=I-B$ and $n\times 1$ vector f, we may construct vectors $x_0, f_0=Qx_0\in\mathbf{R}^2$ and $2n\times 2n$ matrix $\tilde{B}$ such that n-vector x satisfies both the following equations:

$$(I-B)x=f \quad\text{and}\quad (I-\tilde{B})(x_0\otimes x)=f_0\otimes f \tag{5.1}$$

where

$$\tilde{B}=P\otimes I_n+Q\otimes B. \tag{5.2}$$

- By choosing 2×2 matrices Q and P appropriately, the spectra of iteration matrices B and $\tilde{B}$ are related via

$$\sigma(\tilde{B})^2=\varphi(\sigma(B)) \tag{5.3}$$

where φ is any monic quadratic polynomial with the property that $\varphi(1)=1$. (See Theorem 4.1.) Polynomial φ is uniquely determined by its roots $r, s\neq 1$ and therefore has form

$$\varphi(z)=\frac{(z-r)(z-s)}{(1-r)(1-s)}. \tag{5.4}$$

Remark. We assume, henceforth, that matrices Q and P induce the polynomial φ with *real* roots $r\leqq s$ where $r, s\neq 1$. This means that φ is real-valued over the real line. Therefore, whenever the spectrum of $A=(I-B)$ is real, then φ will be real-valued over $\sigma(B)$.

5.1. Spectral assumption. Until now, $\sigma(A)=\sigma(I-B)$ was assumed to be arbitrary except for the fact that $0\notin\sigma(A)$ (equivalently, $1\notin\sigma(B)$). For the first time, we invoke property (1.8), which says that $\sigma(A) = \sigma(I - B)$ is *real and straddles the origin.* That is, we consider the special case

$$(5.5a) \qquad \sigma(A) \subset [a, b] \cup [c, d], \qquad a \leqq b < 0 < c \leqq d.$$

If we set $\alpha = 1 - a$, $\beta = 1 - b$, $\gamma = 1 - c$ and $\delta = 1 - d$, then we obtain the equivalent statement

$$(5.5b) \qquad \sigma(B) \subset [\delta, \gamma] \cup [\beta, \alpha], \qquad \delta \leqq \gamma < 1 < \beta \leqq \alpha.$$

Given the spectral assumptions (5.5a, b), we choose roots r and s so as to tailor $\sigma(\tilde{B})$ according to the linking condition (5.3). This means we can create the optimal configuration for $\sigma(\tilde{B})$ in preparation for a final *two-part splitting acceleration.* The next section provides details.

5.2. Use of two-part splittings. We see directly from (5.3) that values of $\sigma(\tilde{B})$ can be determined by the quadratic polynomial φ of (5.4) which is defined over the intervals (5.5b). Moreover, $\varphi(1) = 1$ is uniquely defined by its roots r, $s \neq 1$.

In the theory of two-part splittings [4], we capture (cover) $\sigma(\tilde{B})$ with the smallest ellipse possible. If this ellipse is symmetric with respect to the real and imaginary axes and has real vertices $\pm H$ and imaginary vertices $\pm iV$ such that

$$(5.6) \qquad 0 \leqq H(r, s) < 1 \quad \text{and} \quad 0 \leqq V(r, s),$$

then a two-part sequence $\{\tilde{y}_k\}$ may be constructed such that convergence to the unique solution vector is guaranteed. Here is how the two-part theory uses the scalars H and V to construct the accelerated sequence $\{\tilde{y}_k\}$.

First, we use the ellipse semi-axes $H(r, s)$ and $V(r, s)$ of (5.6) to define the utility scalar $\rho(r, s)$ where

$$(5.7) \qquad \rho(r, s) = \rho(H, V) = \frac{\sqrt{1 + V^2 - H^2} - 1}{V - H}.$$

The two-part sequence of $2n \times 1$ vectors is defined as follows (see [4]): For arbitrary $2n \times 1$ vectors $\tilde{y}_0$, $\tilde{y}_1$ set

$$(5.8) \qquad \tilde{y}_k = (1 + \Theta\rho^2)\tilde{B}\tilde{y}_{k-1} - \Theta\rho^2\tilde{y}_{k-2} + (1 + \Theta\rho^2)(f_0 \otimes f)$$

for all $k = 0, 1, 2, \cdots$, where $\Theta = (H - V)/(H + V)$.

Scalar ρ of (5.7) describes $R_{\tilde{y}}$, the asymptotic convergence rate, of $\{\tilde{y}_k\} \to x_0 \otimes x$ by the relation

$$(5.9) \qquad R_{\tilde{y}} = -\log(\rho(H, V)).$$

Once we choose the roots r, s for the quadratic φ where $\varphi(1) = 1$, then matrices Q and P may be constructed so that the linking condition (5.3) holds. This is summarized in Table 5.10.

5.3. Minimization. How do we minimize $\rho(H, V)$ in (5.10[iv]), i.e., how do we maximize $R_{\tilde{y}}$ in (5.10[v])? We will see that the answer depends *only* on the interior points b, c of the spectral intervals in (5.5a). In (5.10[i]), choose real roots

$$r = \gamma = 1 - c \quad \text{and} \quad s = \beta = 1 - b.$$

As it turns out, the resulting optimizing iteration matrix $\tilde{B}$ will have pure imaginary spectrum, i.e., $H(r, s) = 0$. (see (5.10[ii], [iii] and Lemma 5.5). Here is the theorem now.

TABLE 5.10
Construction of optimal algorithm via two-part sequence (5.8).

[i]	Select roots $\tau \leqq s \neq 1$ for the quadratic polynomial φ of (5.4) defined over $\sigma(B)$ of (5.5b), which
[ii]	. . . produces iteration matrix $\tilde{B}$ of (5.2) whose spectrum (see (5.3)), in turn,
[iii]	. . . is covered by an ellipse, with semi-axes $H(r, s)$ and $V(r, s)$ (see (5.6)), which
[iv]	. . . produces scalar $\rho(H, V)$ (see (5.7)), which
[v]	. . . defines the final two-part sequence $\{\tilde{y}_k\} \rightarrow (x_0 \otimes x)$ (see (5.8)) with convergence ($R_{\tilde{y}} = -\log(\rho(H, V))$ (see (5.9)).

THEOREM 5.1. *Given the invertible $n \times n$ linear system $Ax = (I - B)x = b$, and the quadratic polynomial*

$$\varphi(z) = \frac{(z-r)(z-s)}{(1-r)(1-s)}, \tag{5.11}$$

which defines the $2n \times 2n$ iteration matrix $\tilde{B}$ of (5.8), *then the maximum convergence rate $R_{\tilde{y}}$ for sequence $\{\tilde{y}_k\}$* (5.8) *occurs when the real roots r, s are chosen such that*

$$r = 1 - c \quad and \quad s = 1 - b, \tag{5.12}$$

i.e., when

$$\varphi(z) = \frac{(z-1+b)(z-1+c)}{bc}, \tag{5.13}$$

in which case the optimal convergence rate is given by

$$R_{\text{opt}} = -\log(\rho_{\text{opt}}) \tag{5.14}$$

where

$$\rho_{\text{opt}} = \frac{\sqrt{1+M^2}-1}{M} \tag{5.15}$$

and

$$M^2 = \max\left\{\frac{(b-a)(a-c)}{bc}, \frac{(b-d)(d-c)}{bc}\right\}. \tag{5.16}$$

The proof develops through the following sequence of lemmas.

Goal. To optimize the convergence of sequence (5.8), we must choose real roots $r \leqq s$ of quadratic polynomial φ which minimizes $\rho(r, s)$ of (5.7). The following sequence of lemmas shows this to happen when roots r and s meet condition (5.12), i.e., when $r = \gamma = 1 - c$ and $s = \beta = 1 - b$.

The first lemma says that if (5.8) converges at all, then the real roots r and s must straddle (or lie on either side of) $z = 1$:

LEMMA 5.2. *Let real roots $r \leqq s$ define quadratic polynomial φ which induces the spectrum $\sigma(\tilde{B})$ as per* (5.3). *If $\sigma(\tilde{B})$ lies in the vertical strip $-1 <$ Real $(z) < 1$, then the real roots $r \leqq s$ straddle the point $z = 1$, i.e.,*

$$r<1<s. \tag{5.17}$$

Proof. Recall from (5.6) that it is *necessary* that $\sigma(\tilde{B})$ lie in the vertical strip $-1 <$ Real $(z) < 1$ if the two-part sequence (5.8) is to converge. Consider the case contrary to (5.17), i.e., either

$$1<r\leqq s \quad \text{or} \quad r\leqq s<1. \tag{5.18}$$

Construction of φ in (5.11) implies that $\varphi(1) = 1$. Therefore, if both real roots lie wholly on one side of $z = 1$, then φ is concave upward over its real domain $\sigma(B)$ (see (5.5b)). This means that for some $\mu_0 \in \sigma(B)$, we have $1 < \varphi(\mu_0)$. Now the linking condition $\sigma(\tilde{B}) = \pm\sqrt{\sigma(B)}$ of (5.3) tells us two things: (i) If $1 < \varphi(\mu_0)$, then $1 < \sqrt{\varphi(\mu_0)} \in \sigma(\tilde{B})$, and (ii) The eigenvalues $\sigma(\tilde{B})$ form a set which is symmetric with respect to the real and imaginary axes. These two conditions together imply that any (necessarily symmetric) covering ellipse for $\sigma(\tilde{B})$ will have real semi-axis $H > 1$, a condition which, in the theory of two-part splitting, guarantees *divergence* of the sequence (5.8). That is, condition (5.18) cannot hold if convergence is to obtain. Therefore, (5.17) is established and the lemma is proved. □

We have just shown that the roots r, s of φ of (5.11) must straddle the point $z = 1$ as per (5.17). We can now state that the crucial values H^2, V^2 of (5.6) and hence, of (5.7), are necessarily among the five values $-\varphi(\alpha)$, $-\varphi(\delta)$, $\varphi(\beta)$, $\varphi(\gamma)$, 0. Here is the lemma that demonstrates this fact.

LEMMA 5.3. *Given quadratic polynomial φ of* (5.11) *subject to condition* (5.17). *If ρ of* (5.7) *is minimal, then the zero-centered ellipse that captures $\sigma(\tilde{B})$ has semi-axes V on the imaginary axis and H on the real axis given by*

$$H^2 = \max\{0, \varphi(\beta), \varphi(\gamma)\}, \tag{5.19a}$$

$$V^2 = \max\{0, -\varphi(\alpha), -\varphi(\delta)\}. \tag{5.19b}$$

Proof. Note that condition (5.17) implies that φ, which is defined over intervals (5.5b), is concave downward. Since $\varphi(1) = 1$, this maximum value is greater than or equal to one. We first argue that the positive maximum of φ *must* occur over the interval (γ, β). If the maximum occurred anywhere outside this interval, then φ would take on a value greater than one over $\sigma(B)$, its domain (5.5b). From (5.3), $\sigma(\tilde{B})$ would then have a real value greater than one. This, in turn, would imply that the smallest covering ellipse for $\sigma(\tilde{B})$ would have a real semi-axis H greater than one, which, from (5.6), is disallowed.

Since the parabola φ over $\sigma(B)$ of (5.5) has its maximum over the open interval (γ, β), its maximum nonnegative value over $\sigma(B)$ occurs at one of the interior end-points β or γ, while its maximum nonpositive value at the outside end-points α or δ. This establishes the lemma. □

The next lemma refines (5.19a) by showing that if $\rho(H, V)$ of (5.7) is minimal, then necessarily, $\varphi(\beta) = \varphi(\gamma)$.

LEMMA 5.4. *Let quadratic polynomial φ of* (5.11) *be chosen so that ρ of* (5.7) *is minimal. Then H^2 is defined by*

$$H^2 = \varphi(\beta) = \varphi(\gamma) \geqq 0. \tag{5.20}$$

Moreover, if $H^2 = 0$, then

$$r = \gamma = 1 - c \quad \textit{and} \quad s = \beta = 1 - b. \tag{5.21}$$

Proof.

Case 1 ($H^2 > 0$). From (5.19a), we may suppose that H^2 is determined by the value of φ at one of the interval end-points, β, say. This implies that root s is to the right of β, (see (5.5b) for $\sigma(B)$ and (5.11) for the definition of φ defined over $\sigma(B)$). That is,

$$\varphi(\beta) = H^2 > 0, \tag{5.22a}$$

$$\varphi(\beta) > \varphi(\gamma), \tag{5.22b}$$

$$\beta \leqq s. \tag{5.22c}$$

Along with constraints (5.22a, b, c) above, we always have as a consequence of (3.8a), the fixed-point property

$$\varphi(1) = 1. \tag{5.23}$$

As roots r and s vary so as to preserve the two fixed point properties, (5.22a) and (5.23), we see that either roots r and s move away from each other or else they move toward each other. As roots r and s move away from each other, the positive values $-\varphi(\alpha)$ and $-\varphi(\delta)$ both decrease and $\varphi(\gamma)$ increases. That is,

$$\begin{array}{l}\text{if } \varphi(1) = 1,\ \varphi(\beta) = H^2,\ r \text{ decreases, } s \text{ increases,}\\ \text{then } -\varphi(\alpha),\ -\varphi(\delta), \text{ both decrease while } \varphi(\gamma) \text{ increases.}\end{array} \tag{5.24}$$

From (5.19b) we see that either $-\varphi(\alpha) = V^2$ or $-\varphi(\delta) = V^2$. Condition (5.24) therefore implies that V^2 decreases as roots r and s separate. At the same time, (5.24) tells us that $\varphi(\gamma)$ increases while H is fixed.

But if V^2 decreases as H is held constant, then ρ decreases. To see this, compute the partials of ρ in (5.7) to obtain

$$\frac{\partial \rho}{\partial H} > 0 \quad \text{and} \quad \frac{\partial \rho}{\partial V} > 0. \tag{5.25}$$

This shows that $\rho(H, V)$ decreases as either H or V decreases.

To minimize ρ, therefore, we seek to move roots r and s as far from each other as possible. How far apart can they be? From (5.22b) and (5.22c), we see that this separation of r and s may increase until $\varphi(\gamma)$ increases to its maximum allowed value, viz., until $\varphi(\gamma) = \varphi(\beta) = H^2$. This proves (5.20) when $H > 0$.

Case 2 ($H^2 = 0$). In this case, it must be that both roots of φ lie inside the closed interval $[\gamma, \beta]$. That is,

$$\gamma \leqq r \quad \text{and} \quad s \leqq \beta. \tag{5.26}$$

Therefore, φ is negative or zero over all of its domain $\sigma(B)$ (see (5.5b), (5.19) and (5.11)) so that $H = 0$ in (5.6). Now as roots r and s separate, i.e., as $r \downarrow \gamma$ and $s \uparrow \beta$, the value of V decreases. Since (5.26) plus (5.19) implies that H is constant and equal to zero, decreasing V implies that ρ of (5.7) decreases also. Thus, ρ is not minimal if r and s are not maximally separated in (5.26). In other words, equality holds everywhere in (5.26). This proves (5.21) for the case $H^2 = 0$ and the lemma is done. □

The next lemma tells us that if ρ is minimal, then equality must obtain everywhere in (5.20).

LEMMA 5.5. *Let quadratic polynomial* φ *of* (5.11) *be chosen so that* ρ *of* (5.7) *is minimal. Then* $H^2 = 0$.

Proof. From (5.20) of Lemma 5.4, we know that $H^2 = \varphi(\beta) = \varphi(\gamma)$. This means that the roots of quadratic φ must be equidistant from the points β and γ. Since φ is concave downward (see (5.17)), the roots lie *outside* the open interval (γ, β). That is, for the single real parameter $t \geqq 0$, we may characterize the root pairs $r = r_t$ and $s = s_t$ by writing

$$r = r_t = \gamma - t \quad \text{and} \quad s = s_t = \beta + t \quad \text{where } t \geqq 0. \tag{5.27}$$

Then the one-parameter family of φ_t (5.11) takes on the form

$$\varphi_t(x) = \frac{(x-t-\beta)(x+t-\gamma)}{(1-t-\beta)(1+t-\gamma)}. \tag{5.28}$$

From (5.20) and (5.19b) respectively, we note that

$$H^2 = H_t^2 = \varphi_t(\beta) = \varphi_t(\gamma),$$

while

$$V^2 = V_t^2 = -\varphi_t(\alpha) \quad \text{or} \quad -\varphi_t(\delta).$$

Differentiate H^2 with respect to t and evaluate at $x = \beta$. (By symmetry of the roots with respect to β and γ, we could just as well evaluate the derivative at $x = \gamma$.) Similarly, differentiate V^2 with respect to t and evaluate at $x = \alpha$ or $x = \delta$. We then obtain

$$2HH' = \frac{d}{dt}H^2 = \frac{d}{dt}\varphi_t(\beta), \tag{5.29a}$$

$$2VV' = \frac{d}{dt}V^2 = -\frac{d}{dt}\varphi_t(\Theta), \qquad \Theta = \alpha, \delta. \tag{5.29b}$$

Through laborious calculation, it follows that the *quotient* of the derivatives (5.29a) and (5.29b) is a negative constant N, which is independent of parameter t. In fact, the exact value of negative N is

$$N \equiv \frac{HH'}{VV'} = \frac{(\beta-1)(\gamma-1)}{(1-\Theta)[\beta+\gamma-(\Theta+1)]}, \qquad \Theta = \alpha, \delta. \tag{5.30}$$

We use N in the following calculation of ρ', the derivative of ρ of (5.7) with respect to t. Accordingly,

$$\rho' = \frac{\partial\rho}{\partial H}\frac{dH}{dt} + \frac{\partial\rho}{\partial V}\frac{dV}{dt}$$

$$= \frac{\partial\rho}{\partial H}H' + \frac{\partial\rho}{\partial V}V'.$$

Multiply this equation by $HV > 0$ to obtain

$$HV\rho' = V\frac{\partial\rho}{\partial H}HH' + H\frac{\partial\rho}{\partial V}VV'.$$

Substitute $HH' = NVV'$ from the LHS of (5.30) in the equation above; then divide by $HV > 0$ to obtain

$$\rho' = \left(\frac{N}{H}\frac{\partial\rho}{\partial H} + \frac{1}{V}\frac{\partial\rho}{\partial V}\right)V'. \tag{5.31}$$

We now show that the sign of ρ' is positive in a neighborhood of $t = 0$ and that as t increases, this sign may change at most once. Now if for all positive t sufficiently close to zero, $r = r(t)$ is arbitrarily close to γ and $s = s(t)$ is arbitrarily close to β which implies that H is arbitrarily close to zero. This means that N/H in (5.31) is arbitrarily large and negative. At the same time, from (5.25), $(\partial\rho/\partial H) > 0$ and $(\partial\rho/\partial V) > 0$. Finally, (5.24) has it that $V' < 0$ so that we may conclude from (5.31) that $\rho' > 0$ over some neighborhood $0 \leqq t \leqq N_0$.

Now as t increases, the derivative ρ' remains positive and changes sign at most once. Note that V decreases as t increases while H increases. That is, the absolute value of the negative term N/H of (5.31) decreases while that of the positive term $1/V$ increases. As t increases, therefore, it is possible (not guaranteed) that the RHS of (5.31) will change from positive to negative. In any event, a minimum value of ρ is realized when $r = \gamma$ and $s = \beta$. From (5.20), this guarantees that $H^2 = 0$. This proves the lemma. □

Remark. In theory, we could find ρ_{opt} in (5.15) by showing that $d\rho/dt > 0$ for all $t \geqq 0$. First, write $\rho(t)$ as a function of t using (5.7), (5.27) and (5.28). Secondly, compute the derivative of ρ with respect to t. The derivative $d\rho/dt$ was, indeed, computed using MACSYMA, and the symbolic result for this derivative consisted of some *two hundred* lines. This voluminous output, however impressive and accurate, refused to reveal the sign of $\rho'(t)$. The lemmas above, therefore, serve as an alternative to the MACSYMA result.

All the pieces are in place for the proof of Theorem 5.1.

Proof of Theorem 5.1. Lemma 5.5 assures us that $H^2 = 0$, which, from (5.21), says that $r = \gamma = 1 - c$ and $s = \beta = 1 - b$. Substituting these values for r and s into (5.11) produces the form (5.13) of Theorem 5.1. The expression for M^2 given by (5.16) is just a restatement of (5.19b) after setting $M = V$. Now ρ as expressed by (5.15) is obtained from the definition (5.7) where $H = 0$ and $V = M$. Finally, convergence rate R_{opt} of (5.14) is always given in terms of ρ, as expressed in (5.14) (see e.g., [4], [23]). All the conditions of Theorem 5.1 have been justified, so Theorem 5.1 is proved. □

6. The basic algorithm.

We now present a form of pseudocode that details the implementation of the algorithm (see Table 6.1). The justification appears at the end of this section.

TABLE 6.1
Optimal algorithm.

OBJECTIVE:	To find solution vector x for the linear system $Ax = f$.
INPUT:	
A	invertible $n \times n$ matrix,
$a \leqq b < 0 < c \leqq d$	spectral parameters for matrix A, $\sigma(A) \subset [a, b] \cup [c, d]$,
f	the n-vector output,
tol	positive cut-off tolerance,
k_max	maximum iteration count,
v_0, v_1, w_0, w_1	arbitrary initial n-vectors.
OUTPUT:	
$x_0 = [x_{01}, x_{02}]$	auxiliary 2-vector
v_k, w_k	auxiliary n-vectors, $k = 2, 3, \cdots$
step_count	expected step count: estimated number of iterations which suffice to produce each additional decimal place of accuracy,
$x := (v_k/x_{01})$	if $\| Av_k - x_{01} f \| < tol$ and $k < k$_max,
$x := (w_k/x_{02})$	if $\| Aw_k - x_{02} f \| < tol$ and $k < k$_max.

Step 1: Compute 2×2 matrices Q, P as follows:

$$\text{Set } r := 1 - c, \tag{6.2}$$

$$s := 1 - b, \tag{6.3}$$

$$p := \sqrt{\frac{rs}{(1-r)(1-s)}}, \tag{6.4}$$

$$q := -\frac{p(r+s)}{2rs}, \tag{6.5}$$

$$q_{12} := \sqrt{|1-(p+q)^2|}, \tag{6.6}$$

$$q_{21} := (1-(p+q)^2)/q_{12}, \tag{6.7}$$

$$Q := \begin{bmatrix} q & q_{12} \\ q_{21} & -q \end{bmatrix}, \tag{6.8}$$

$$P := \begin{bmatrix} p & 0 \\ 0 & -p \end{bmatrix}. \tag{6.9}$$

Step 2: Compute auxiliary 2×1 vectors $x_0 = [x_{01}, x_{02}]$, and $f_0 = [f_{01}, f_{02}]$, as follows:

$$x_0 := \text{null vector of } (Q+P-I) \tag{6.10}$$

$$\text{i.e., } (Q+P-I)x_0 = 0.$$

$$f_0 := Qx_0. \tag{6.11}$$

Step 3: Compute constants M and ρ where

$$M := \left[\max \left\{ \frac{(b-a)(a-c)}{bc}, \frac{(b-d)(d-c)}{bc} \right\} \right]^{1/2}, \tag{6.12}$$

$$\rho := \frac{\sqrt{1+M^2}-1}{M}. \tag{6.13}$$

Step 4: Return the expected step count.

$$step\_count := \frac{-1}{\log 10(\rho)}. \tag{6.14}$$

Step 5: Compute iteratively the sequence of n-vector pairs $\{v_k, w_k\}$, $k = 2, 3, 4, 5, \cdots$.

REPEAT $k = 2, 3, 4, 5, \cdots$,

$$\begin{aligned} v_k := {} & (1-\rho^2)[-A(qv_{k-1} + q_{12}w_{k-1}) + (q+p)v_{k-1} + q_{12}w_{k-1}] \\ & + \rho^2 v_{k-2} + (1-\rho^2) f_{01} f, \end{aligned} \tag{6.15}$$

$$\begin{aligned} w_k := {} & (1-\rho^2)[A(qw_{k-1} - q_{21}v_{k-1}) - (q+p)w_{k-1} + q_{21}v_{k-1}] \\ & + \rho^2 w_{k-2} + (1-\rho^2) f_{02} f, \end{aligned} \tag{6.16}$$

$$\text{UNTIL } \|Av_k - x_{01} f\| < tol \tag{6.17a}$$

$$\text{or } \|Aw_k - x_{02} f\| < tol \tag{6.17b}$$

$$\text{or } k \geqq k\_\max. \tag{6.17c}$$

Step 6: If $k < k\_\max$, then return solution vector x, viz.,

(6.18a) $$x := v_k/x_{01} \quad \text{if (6.17a) holds,}$$

(6.18b) $$x := w_k/x_{02} \quad \text{if (6.17b) holds.}$$

Table 6.1′ presents an overview of this algorithm along with references to the particular displays that justify Steps 1–6.

The justifying displays above which correspond to Step 1, Step 2, Step 3 and Step 4 are fairly straightforward. Step 5 and Step 6, however, are more fully explained as follows.

Step 5: To validate the simultaneous iterative steps (6.14) and (6.15), note that Lemma 5.5 has it that $H = 0$. We conclude from (5.8) that $\Theta = -1$. Thus, the two-part sequence for $\{\tilde{y}_k\}$ in (5.8) assumes the form

$$\tilde{y}_k = (1-\rho^2)\tilde{B}\tilde{y}_{k-1} + \rho^2\tilde{y}_{k-2} + (1-\rho^2)(f_0 \otimes f),$$

which, from (5.2), gives us

(6.19) $$\tilde{y}_k = (1-\rho^2)(P \otimes I + Q \otimes B)\tilde{y}_{k-1} + \rho^2\tilde{y}_{k-2} + (1-\rho^2)(f_0 \otimes f),$$

for all $k = 0, 1, 2, \cdots$. Now the sequence (6.19) above consists of $2n$-vectors $\{\tilde{y}_k\}$ which converges to the tensor vector $(x_0 \otimes x)$, where $Ax = f$; see (5.1). That is,

(6.20) $$\tilde{y}_k = \begin{bmatrix} v_k \\ w_k \end{bmatrix} \rightarrow x_0 \otimes x = \begin{bmatrix} x_{01}x \\ x_{01}x \end{bmatrix} \quad \text{as } k \uparrow \infty.$$

Note that (6.20) defines each $2n$-vector $\tilde{y}_k$ in terms of the n-vectors v_k and w_k. Now substitute the left-hand equality of (6.20) into (6.19) and use the tensor definition (2.1) to obtain

(6.21) $$\begin{aligned} \begin{bmatrix} v_k \\ w_k \end{bmatrix} &= (1-\rho^2) \cdot \begin{bmatrix} pI_n & 0 \\ 0 & -pI_n \end{bmatrix} \begin{bmatrix} v_{k-1} \\ w_{k-1} \end{bmatrix} \\ &+ \begin{bmatrix} qB & q_{12}B \\ q_{21}B & -qB \end{bmatrix} \begin{bmatrix} v_{k-2} \\ w_{k-2} \end{bmatrix} + \begin{bmatrix} \rho^2 v_{k-2} \\ \rho^2 w_{k-2} \end{bmatrix} + (1-\rho^2) \cdot \begin{bmatrix} f_{01}f \\ f_{02}f \end{bmatrix}. \end{aligned}$$

TABLE 6.1′
Optimal algorithm.

	Condition	Justification
Step 1:	(6.2), (6.3) (6.4) (6.5) (6.6), (6.7) (6.8), (6.9)	(5.12) (4.5a) (4.5b) (4.5c) (4.4)
Step 2:	(6.10) (6.11)	(3.8a, b) (3.11)
Step 3:	(6.12) (6.13)	(5.16) (5.15)
Step 4:	(6.14)	(1.11)
Step 5:	(6.15), (6.16), (6.17)	see below
Step 6:	(6.18)	see below

If we replace B with $(I_n - A)$ in (6.21) then block matrix multiplication produces (6.14) and (6.15), respectively. From (6.20), we see that $v_k \rightarrow x_{01}x$ and $w_k \rightarrow x_{02}x$. Since $Ax = f$, we see that the stopping criteria, (6.17a, b, c) of Step 5, are justified.

Step 6: We see from (6.20), that solution vector $x = \lim_{k\to\infty} v_k/x_{01}$ and $x = \lim_{k\to\infty} w_k/x_{02}$. This validates (6.18a) and (6.18b) of Step 6 *if* the tolerance conditions (6.17a) and (6.17b) are met.

7. Concluding remarks. We present some observations on the algorithm of the previous section which concern possible shortcuts, the condition number, avoidance of complex arithmetic, and some numerical results.

Short cuts.

- Simplifying x_0: Note that the vector x_0 in (6.10) is a nonzero null vector of the 2×2 matrix $Q + P - I$. Its computation, therefore, involves a separate subroutine. Now we may always set one of the entries of x_0 to one, say $x_{01} = 1$. Then steps (6.18a) and (6.18b) are simplified in the sense that x_0 does not have to be "factored out" in order to return solution vector x.
- A New Test Vector $z_x(\lambda)$: We have observed experimentally that on occasion, a barycentric linear combination

$$z_k(\lambda) = \lambda \frac{v_k}{x_{01}} + (1-\lambda)\frac{w_k}{x_{02}}, \qquad \lambda \in \mathbf{R} \tag{7.1}$$

will be closer to the solution vector x than either of the candidate vectors $\{v_k/x_{01}\}$ or $\{w_k/x_{02}\}$. (See (6.18a) and (6.18b).) That is, it may happen that for a certain scalar λ_0,

$$\|Az_{k_0}(\lambda_0) - f\| < tol \tag{7.2}$$

for some k_0 that occurs much *earlier* than the k that we would find in (6.18a) or (6.18b). During execution, then, periodic testing of vector $z_k(\lambda)$ for various values of λ might accelerate the convergence of the algorithm. In fact, if (7.2) obtains for some k_0, and (6.18a) and (6.18b) both fail, then overwrite all current vectors $v_{k-1}, w_{k-1}, v_{k-2}, w_{k-2}$ with the closer approximating vector (7.1).

Condition number. Experimental results[1] reveal that the condition number of our original system (1.1) and of our tensor equivalent system (1.2), (1.3) are of the same order of magnitude. That is, the condition number does not increase. One example we tested was the system $Ax = f$, where A was taken to be a small 12×12 Hermitian matrix constructed from the notoriously badly-conditioned Hilbert matrix. The spectrum of A lay on both sides of the origin. We applied our algorithm to $Ax = f$ and we used the conjugate gradient method on the positive definite normal system $A^*Ax = A^*f$. Even though the conjugate gradient method held the advantage of a small dimension (in exact arithmetic, the conjugate gradient method converges after 12 steps), the condition number was high enough to render this method unstable. Our own method applied to $Ax = f$ behaved as predicted by *step_count* of (6.14).

The linking condition (1.5a), (1.5b) allows arbitrary quadratic polynomials $\varphi(1)$ to be used. Preliminary results now indicate that, in some special cases, the condition number of the equivalent tensor system may be *less* (better) than that of the original system.

One more point on condition number: In (6.6) and (6.7) we construct matrix entries q_{12} and q_{21}, which have the same modulus. In fact, (4.5c) only requires the *product* $q_{12}q_{21}$

[1] The PC-MATLAB software was extensively used for generating examples and condition number computations. This software proved invaluable in this research.

to equal $1 - (p + q)^2$; the individual factors are not constrained. It is our empirical experience that equal moduli on the factors q_{12} and q_{21} seem to improve the condition or the stability of our algorithm.

Complex arithmetic. An equivalent expression for scalar p in display (6.4) is

$$p := \sqrt{\frac{(1-b)(1-c)}{bc}}.$$

Since $b < 0$ and $c > 0$ (see (5.5a)), scalar p will be complex if and only if $c < 1$. To guarantee that p is real, then, replace A and f in (1.1) with tA and tf, respectively, where $t > 1/c$. Finally, note that only in (6.4) can complex scalars be created: in (6.5), (6.6) and (6.7), real scalar input results in real output.

Maximum iteration count. One of our input parameters is *k*_max, which denotes the maximum tolerable iteration count. It may be convenient, however, to calculate *k*_max from the computed value (6.14) for *step_count*, the expected number of iterations for each decimal place of accuracy. For example, to assure six-place accuracy, the user may wish to set *k*_max := (6* *step_count*).

Numerical results. Empirical observation suggests that the *step_count* estimate (6.14) is fairly accurate. Recall from (1.11) that *step_count* is $1/R_y$, the reciprocal of the asymptotic rate of convergence. Accordingly, we compare R_y, our rate of convergence, to $R_{(E/N)}$, the convergence rate of Eiermann and Niethammer [6], which treats the case

$$\sigma(A) \subset [-1, -\epsilon] \cup [\epsilon, 1]. \tag{7.3}$$

The resulting asymptotic convergence rate (and corresponding *step_count*) is

$$R_{(E/N)} = \frac{1}{step\_count_{(E/N)}} = -\frac{1}{2}\log\left[\frac{1-\epsilon^2}{1+\epsilon^2}\right]. \tag{7.4}$$

From our theory, the convergence rate R_y for (7.4) is

$$R_y = \frac{1}{step\_count_y} = -\frac{1}{2}\log\left[\frac{1-\epsilon}{1+\epsilon}\right].$$

The Laurent expansion (using MACSYMA) of the ratio of these two convergence rates is

$$\frac{step\_count_{(E/N)}}{step\_count_y} = \frac{R_y}{R_{(E/N)}} = \frac{1}{\epsilon} + \frac{\epsilon}{3} - \frac{2\epsilon^3}{15} + \frac{2\epsilon^5}{63} + \cdots,$$

which shows that $(R_y/R_{(E/N)}) \rightarrow \infty$ as ϵ goes to zero, i.e., as the condition number of A increases.

8. Related work.

Known iterative methods exist for solving (1.1), e.g., Gauss–Seidel, conjugate gradient, successive overrelaxation. But for the most part, these methods require either that the spectrum $\sigma(A)$ lies wholly within an open hyperplane having $z = 0$ on the boundary, or that A possesses additional structure such as symmetry or normality. Here are some of these methods.

8.1. Algorithms for which $\sigma(A)$ cannot straddle the origin.

The successive overrelaxation (SOR) method considers $Ax = (I - B)x = f$ where iteration matrix $B = L + U$. In the terminology of (1.2), the SOR method studies the parameterized class of *equivalent systems* as described in (1.6). The Gauss–Seidel method results when $\omega = 1$. In theory, then, the eigenvalues of A may be well scattered about the complex plane. But practical *a priori* calculation of optimal parameters is possible only for certain special cases—one of these occurs when A is consistently ordered in which case we assume that the spectrum of $B = L + U$ lies in the open real interval $(-1, 1)$; cf., [22], [23], [24]. This implies that $\sigma(A)$ may not straddle the origin.

The early work in semi-iterative methods by Gene Golub and Richard Varga [11], develops a nonstationary method (constants are updated at each iterative step) but the spectrum $\sigma(A)$ is presumed to lie within the real interval (a, b), where $0 \not\in (a, b)$.

Tom Manteuffel [12] presents a nonstationary Chebyshev semi-iterative method for the class of matrices A where $\sigma(A)$ is contained inside an ellipse which, in turn, excludes the origin.

A stationary method in the form of k-part splitting is given by de Pillis in [4], where $\sigma(A)$, the spectrum of matrix A, must be contained within an ellipse—moreover, zero must lie outside this ellipse.

The assumption of diagonal dominance for A which guarantees success of Gauss–Seidel and Jacobi iteration for irreducible matrices, also forbids $\sigma(A)$ to lie on both sides of the origin.

8.2. Algorithms for which $\sigma(A)$ straddles the origin. As we have noted in the introduction, system (1.1) includes all symmetric (Hermitian) systems. The following special case arises in several contexts:

$$\begin{bmatrix} H & C^* \\ C & 0 \end{bmatrix} \quad \text{where } H = H^*. \tag{8.1}$$

(Symbol H^* indicates conjugate transpose of matrix H.)

For example, in T. Markham, M. Neumann and R. Plemmons [13] a large scale linear least squares problem is studied for the $(m + n \times n)$ over-determined system

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \cdot \begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} b \\ f_1 \end{bmatrix} \tag{8.2}$$

where $n \times n$ matrix A_1 is invertible. The authors show that (8.2) is equivalent to the linear system

$$\begin{bmatrix} A_1 & 0 & I_m \\ A_2 & I_n & 0 \\ 0 & A_2^* & A_1^* \end{bmatrix} \cdot \begin{bmatrix} y \\ r_2 \\ r_1 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix}. \tag{8.3}$$

If we interchange the first and third block columns of (8.3), then we see that the least squares problem can be cast in form (8.1) with $H = I_{m+n}$.

In a very recent work of Plemmons [19], the minimization problem

$$\tfrac{1}{2}\langle Ax, x\rangle - \langle s, x\rangle \quad \text{where } Ex = t \tag{8.4}$$

is considered where matrix A is positive semidefinite. Once again, system (8.3) is converted to the equivalent linear system

$$\begin{bmatrix} A & E^* \\ E & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix}, \tag{8.5}$$

which is of form (8.1).

Iterative methods for solving (8.5) appear in Dyn and Ferguson [5], where various splittings are used. Conjugate gradient techniques can be found in Axelsson's work [1].

The work of Fix, Gunzburger and Nicolaides [10] discusses mixed finite element schemes. (Mixed schemes are finite element approximations based on stationary variational principles instead of strict maxima or minima.) These authors develop their work by viewing the mixed method as abstract Galerkin methods applied specifically to operator equations of the form (8.1).

Recently, Wilhelm Niethammer and Richard Varga [15] presented a k-step stationary method in which each vector iterate v_n, $n > k$, depends on the k previous vectors as indicated by (3.12). In this work, *any* compact configuration for $\sigma(A)$, straddling or not, can (theoretically) be accommodated for given invertible A if k is taken sufficiently large. But calculation of the parameters in general is not yet practical. The case $k = 2$, however, is well understood—$\sigma(A)$ may lie within any ellipse with $z = 0$ in it exterior. The case of straight line spectrum for A is obtained when the ellipse is degenerate. But when $k = 2$, zero is in the exterior of any allowable ellipse—the origin may not be straddled.

The conjugate gradient method traditionally treats the case A is positive definite Hermitian. Hence, $\sigma(A)$ can not straddle the origin. However, Concus and Golub [2] present a nonstationary variation which can deal with matrices of form (8.5). In a more recent work, V. Faber and T. Manteuffel [9] extended this technique to the wider class of matrices A which are either (i) Hermitian or (ii) a linear combination of the identity and a skew symmetric $S = -S^*$. Thus, $\sigma(A)$ may straddle the origin but matrix A must always be normal ($AA^* = A^*A$). Also, the conjugate gradient method is nonstationary.

A very general class of matrices is treated by C. Paige and M. Saunders [17]. Here, solution of (1.1) is regarded as a minimization problem in a least squares setting which uses the Golub–Kahan lower bi-diagonal reduction of A. This method is nonstationary. Also, in the work of Opfer and Schober [16], the general $n \times n$ matrix A is treated by constructing a polynomial with a specified min-max property. One interesting result (Corollary 5.1) is this: if the iteration scheme is to be *stationary*, then the theory guarantees existence of an optimal polynomial for matrix A if and only if $\sigma(A)$ does not straddle the origin. (Our paper presents a stationary scheme for straight-line spectra which *do* straddle the origin.)

Solution to (1.2) is also given by Y. Saad [20] for the zero-straddling case $\sigma(A) \subset [a, b] \cup [c, d]$ where $a < b < 0 < c < d$. This iterative technique is reminiscent of the Chebyshev semi-iterative methods except that least square polynomials are used in place of the Chebyshev minimax polynomials. Note that A must be self-adjoint. Similarly, in the work of de Boor and Rice [3] an optimal Chebyshev type of polynomial is developed under the assumption that matrix A is real and symmetric. (Our paper assumes a straight line spectrum for A but requires neither the real or symmetric properties.)

In a similar vein, D. Smolarski [21] deals with "boomerang shaped" spectra. This is, $\sigma(A)$ lies in a polygon which is symmetric with respect to the real axis. This method is nonstationary.

In the work by M. Eiermann and W. Niethammer [6] Euler summation theory is the cornerstone for treating $\sigma(A)$ of arbitrary bounded shape. A nonstationary algorithm is developed for straight line zero-straddling $\sigma(A)$ in $[a, b] \cup [-b, -a]$ where $a < b < 0$. (See (7.4).)

A comprehensive overview of iterative methods may be found in the paper of Howard Elman [8]. See also [18] by Patterson.

Acknowledgments. Many thanks to Professors Al Kelley, University of California, Santa Cruz and Fred T. Metcalf, University of California, Riverside, for many helpful discussions regarding MACSYMA.

REFERENCES

[1] O. AXELSSON, *Numerical algorithms for indefinite problems*, in Elliptic Problem Solvers, II, Academic Press, New York, 1985.

[2] P. CONCUS AND G. GOLUB, *A generalized conjugate gradient method for non-symmetric systems of linear equations*, Proc. Second International Symposium on Computing Methods in Applied Sciences and Engineering, December 1975.

[3] C. DE BOOR AND J. R. RICE, *Extremal polynomials with applications to Richardson iteration for indefinite linear systems*, SIAM J. Sci. Statist. Comput, 3 (1982), pp. 47–57.

[4] J. DE PILLIS, *How to embrace your spectrum for faster iterative results*, Linear Algebra Appl., 34 (1980), pp. 125–143.

[5] N. DYN AND W. FERGUSON, *The numerical solution of equality-constrained quadratic programming problems*, Math. Comp., 41 (1983), pp. 165–170.

[6] M. EIERMANN AND W. NIETHAMMER, *On the construction of semi-iterative methods*, SIAM J. Numer. Anal., 20 (1983), pp. 1153–1160.

[7] M. EIERMANN, W. NIETHAMMER AND R. VARGA, *A study of semi-iterative methods for nonsymmetric systems of linear equations*, 1985.

[8] H. ELMAN, *Iterative methods for large, sparse, nonsymmetric systems of linear equations*, Yale Univ. Dept. of Computer Science, Research Report No. 229, New Haven, CT, 1982.

[9] V. FABER AND T. MANTEUFFEL, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[10] G. J. FIX, M. D. GUNZBURGER AND R. A. NICOLAIDES, *Theory and applications of mixed finite element methods*, in Constructive Approaches to Mathematical Models, Academic Press, New York, 1979.

[11] G. GOLUB AND R. VARGA, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods and second order Richardson iterative methods, Parts* I, II, Numer. Math., 3 (1961), pp. 147–168.

[12] T. MANTEUFFEL, *The Tchebychev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307–327.

[13] T. MARKHAM, M. NEUMANN AND R. PLEMMONS, *Convergence of a direct iterative method for large-scale least squares problems*, Linear Algebra Appl., 65 (1985), pp. 155–167.

[14] W. NIETHAMMER, *On different splittings and the associated iteration methods*, SIAM J. Numer. Anal., 16 (1979), pp. 186–200.

[15] W. NIETHAMMER AND R. VARGA, *The analysis of k-step iterative methods for linear systems from summability theory*, Numer. Math., 41 (1983), pp. 177–206.

[16] G. OPFER AND G. SCHOBER, *Richardson's iteration for nonsymmetric matrices*, Linear Algebra Appl., 58 (1984), pp. 343–361.

[17] C. C. PAIGE AND M. A. SAUNDERS, *An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1) (1982), pp. 43–71.

[18] W. PATTERSON III, *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space—A Survey*, Springer–Verlag, New York, 1974.

[19] R. PLEMMONS, *A parallel block iterative scheme applied to computations in structural analysis*, this Journal 7 (1986), pp. 337–347.

[20] Y. SAAD, *Iterative solution of indefinite symmetric linear systems by methods using orthogonal polynomials over two disjoint intervals*, SIAM J. Numer. Anal., 20 (1983), pp. 784–811.

[21] D. C. SMOLARSKI, *Optimum semi-iterative methods for the solution of any linear algebraic system with a square matrix*, thesis, Univ. of Illinois Dept. of Computer Science, Urbana, IL, 1982.

[22] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[23] R. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[24] D. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

002

INHERITED MATRIX ENTRIES: PRINCIPAL SUBMATRICES OF THE INVERSE*

WAYNE W. BARRETT†, CHARLES R. JOHNSON‡, D. D. OLESKY§ AND P. VAN DEN DRIESSCHE¶

Abstract. For a nonsingular n-by-n matrix $A = [a_{ij}]$, let $\alpha \subseteq \{1, 2, \cdots, n\}$ and let $A[\alpha]$ denote the principal submatrix of A lying in the rows and columns indicated by α. We determine the combinatorial circumstances under which the (i, j) entry of the Schur complement $(A^{-1}[\alpha])^{-1}$ equals a_{ij}, and under which the graph of this Schur complement is contained in the graph of $A[\alpha]$.

Key words. digraph, Gaussian elimination, inverse, principal submatrix, Schur complement, sparse matrix

AMS(MOS) subject classifications. 05C50, 15A09, 65F50, 68R10

1. Introduction. Let $A = [a_{ij}]$ be an n-by-n nonsingular matrix and let $N = \{1, 2, \cdots, n\}$. For index sets α, $\beta \subseteq N$ we denote the submatrix of A lying in rows α and columns β by $A[\alpha|\beta]$; in case $\beta = \alpha$, the submatrix is principal and we abbreviate this to $A[\alpha]$. The set $N - \alpha$ is denoted by α^c. We index the entries of $A[\alpha]$ with their indices from α, so that each entry retains the indices associated with its position in A.

We are interested in questions of the following qualitative type. When do certain entries in a matrix derived from A coincide identically with the corresponding entries of A, or when is the zero pattern in A preserved in a matrix derived from A? For example, provided the inverses exist, it is a familiar fact that

$$(A^{-1}[\alpha])^{-1} = A[\alpha], \quad \text{or equivalently } A^{-1}[\alpha] = (A[\alpha])^{-1}, \tag{1.1}$$

if A is triangular and α is a *consecutive* set of indices; i.e., inversion may be carried out "locally." The matrix $(A^{-1}[\alpha])^{-1}$ is the Schur complement of $A[\alpha^c]$ in A (see e.g., [3], [11]) and arises naturally in Gaussian elimination. It is obvious that (1.1) is valid for *any* nonempty index set α if A is a nonsingular diagonal matrix. If A is a restricted type of upper triangular matrix, namely

$$A = \begin{bmatrix} D_1 & A_{12} \\ 0 & D_2 \end{bmatrix} \tag{1.2}$$

where D_1 and D_2 are nonempty nonsingular diagonal matrices and A_{12} is an arbitrary rectangular matrix, then (1.1) still holds for all α. If A is tridiagonal, then (1.1) does not in general hold; however, for α a consecutive set of indices, the graph of $(A^{-1}[\alpha])^{-1}$ is

* Received by the editors June 16, 1986; accepted for publication September 24, 1986.

† Department of Mathematics, Brigham Young University, Provo, Utah 84602. The work of this author was partially supported by Brigham Young University College of Physical and Mathematical Sciences Summer Research Fellowships.

‡ Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29634-1907. The work of this author was partially supported by National Science Foundation grant DMS8500372 and by the Office of Naval Research contract N00014-86-K-0012. Some of the research was carried out while he was a Lansdowne Lecturer at the University of Victoria, Victoria, British Columbia V8W 2Y2 Canada.

§ Department of Computer Science, University of Victoria, Victoria, British Columbia V8W 2Y2 Canada. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8214.

¶ Department of Mathematics, University of Victoria, Victoria, British Columbia V8W 2Y2 Canada. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8965.

contained in the graph of $A[\alpha]$. This follows from the often noted fact that if A is tridiagonal and nonsingular, then so is the inverse of any nonsingular principal submatrix of A^{-1} (see, e.g., [1, Cor. 3.3]). So in this case there is preservation of the zero pattern rather than particular matrix entries. Our results here unify and generalize these familiar facts by determining the most general combinatorial circumstances under which there is entry preservation or graph containment.

We now state a list of general questions to be addressed, and then give some more notation and definitions which we need.

(QI) Under what combinatorial circumstances is the i, j entry of $(A^{-1}[\alpha])^{-1}$ equal to a_{ij}
(a) For a given α and particular $i, j \in \alpha$?
(b) For a given α and all $i, j \in \alpha$?
(c) For given i, j and all α with $i, j \in \alpha$?
(d) For all α and all $i, j \in \alpha$?

(QII) Under what combinatorial circumstances is the graph of $(A^{-1}[\alpha])^{-1}$ contained in that of $A[\alpha]$
(a) For a given α?
(b) For all α?
(c) For a given α, assuming A is combinatorially symmetric with all $a_{kk} \neq 0$?
(d) For all α, assuming A is combinatorially symmetric with all $a_{kk} \neq 0$?

We consider both the directed and undirected graphs of A. Given a matrix A, its directed graph, $D(A)$, has node set N and a directed edge (i, j) from i to j iff $a_{ij} \neq 0$. Given any directed graph, we say that a matrix B is *consistent* with that graph if $b_{ij} = 0$ whenever there is no edge between i and j in the graph. (Note that b_{ij} may be zero when such an edge exists.)

As our questions are combinatorial in nature, we neglect the possibility of accidental cancellations (see, e.g., [2]). Given a directed graph D, we say that two numbers f and g computable from the entries of a matrix consistent with D are *equal generically* (written $f = g$ (generically)) if $f(A) = g(A)$ for *all* A consistent with D. With D given and for $i, j \in \alpha \subseteq N$, we say that j is *reachable from* i *through* α^c if there exists a path of length $\geqq 2$ in D, say $i \rightarrow p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_k \rightarrow j$, in which all the intermediate nodes $p_1, \cdots, p_k \in \alpha^c$ and are distinct. This is similar to the definition in [8] but there a path can have length 1, i.e., be just the edge (i, j). From our definition, a path has length $\leqq |\alpha^c| + 1$, and in case $i = j$ it is a cycle.

When A is combinatorially symmetric ($a_{ij} \neq 0$ iff $a_{ji} \neq 0$) we also work with the undirected graph, $G(A)$, which has node set N and an undirected edge (i, j) between i and j iff $a_{ij} \neq 0$. In this case we assume that all a_{kk} are nonzero. Given an undirected graph G the definitions of "A is consistent with G" and "$f = g$ (generically)" are analogous to those for the directed graph. In questions (QII) the graph is directed in (a), (b), and undirected in (c), (d). The graph of $A[\alpha]$ is the subgraph of A generated by the nodes in α; thus, the graph of $A[\alpha]$ is contained in the graph of A for all α.

We now state and prove our main results, which enable us in § 2 to answer questions (QI) and to make some observations concerning the relation of our results to Gaussian elimination for sparse matrices. Then, in § 3, we give graphical interpretations to the results to answer questions (QII). Finally, in § 4, we give examples to illustrate our results; a reader might like to consult these during the course of reading the next two sections.

2. Submatrix results. We begin with a relationship between minors of A and of the Schur complement of $A[\alpha^c]$ in A. We abbreviate the determinant of A to $\det A$, and if $\alpha = \emptyset$, then $\det A[\alpha] = 1$.

THEOREM 2.1. *Let A be an n-by-n nonsingular matrix and suppose that $\alpha \subseteq N$ is an index set such that $A[\alpha^c]$ is nonsingular. Then for index sets $I, J \subseteq \alpha$ with $|I| = |J|$, we have*

$$\det [(A^{-1}[\alpha])^{-1}][I|J] = (-1)^s \frac{\det A[\alpha^c \cup I | \alpha^c \cup J]}{\det A[\alpha^c]}$$

where $s = \sum_{i \in I}(r_i + i) + \sum_{j \in J}(c_j + j)$ and $r_i \langle c_j \rangle$ is the position of row i $\langle$column $j\rangle$ of A in $A[\alpha]$.

Proof. By repeated use of Jacobi's identity (see, e.g., [11, p. 21]) we have, for $s_1 = \sum_{i \in I} r_i + \sum_{j \in J} c_j$, $s_2 = \sum_{i \in I} i + \sum_{j \in J} j$, and $s = s_1 + s_2$:

$$\begin{aligned}\det [(A^{-1}[\alpha])^{-1}][I|J] &= \frac{(-1)^{s_1} \det A^{-1}[\alpha - J | \alpha - I]}{\det A^{-1}[\alpha]} \\ &= \frac{(-1)^{s_1}(-1)^{s_2} \det A[\alpha^c \cup I | \alpha^c \cup J]}{\det A \det A^{-1}[\alpha]} \\ &= (-1)^s \frac{\det A[\alpha^c \cup I | \alpha^c \cup J]}{\det A[\alpha^c]}. \qquad \square\end{aligned}$$

Note that if I and J have cardinality one, i.e., $I = \{i\}$ and $J = \{j\}$, with $i, j \in \alpha$, Theorem 2.1 specializes to a formula for the i, j entry of $(A^{-1}[\alpha])^{-1}$ in terms of minors of A.

COROLLARY 2.2. *Let A be an n-by-n nonsingular matrix and suppose that $\alpha \subseteq N$ is an index set such that $A[\alpha^c]$ is nonsingular. Then, for $i, j \in \alpha$ with $s = r_i + i + c_j + j$ we have*

$$(A^{-1}[\alpha])^{-1}_{ij} = (-1)^s \frac{\det A[\alpha^c \cup \{i\} | \alpha^c \cup \{j\}]}{\det A[\alpha^c]}. \qquad \square$$

We note that this gives an expression for the i, j entry of the Schur complement, which is also the i, j entry of the reduced matrix of Gaussian elimination obtained after eliminating on the rows specified by α^c (provided $A[\alpha^c]$ has an LU factorization). The results of Theorem 2.1 and Corollary 2.2 are also given in [4] and [7, p. 26], respectively, for the case $\alpha^c = \{1, 2, \cdots, p\}$.

To facilitate statements, we introduce the following hypothesis which specifies a class of matrices to which our results apply.

(H) Let $\alpha \subseteq N$, let D be a given directed graph on the node set N, let $A = [a_{ij}]$ be *any* nonsingular matrix consistent with D and with $A[\alpha^c]$ nonsingular.

We now state our main result, which provides a necessary and sufficient condition to answer (QI)(a). Although our question is about matrices, our characterization uses graph-theoretic ideas.

THEOREM 2.3. *Assuming* (H), *given α and particular $i, j \in \alpha$, then we have*

$$(A^{-1}[\alpha])^{-1}_{ij} = a_{ij} \qquad (\textit{generically}) \tag{2.1}$$

iff either

(i) *j is not reachable from i through $\alpha^c \equiv N - \alpha$, or*

(ii) *if j is reachable from i through vertices $p_1, p_2, \cdots, p_t \in \alpha^c$, then*

$$\det A[\alpha^c - \{p_1, \cdots, p_t\}] = 0 \qquad (\textit{generically}).$$

(*Note that the "if" implication still holds if the equalities are not generic.*)

Proof. From Corollary 2.2, we have $(A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$ iff

$$\det A[\alpha^c \cup \{i\} | \alpha^c \cup \{j\}] = (-1)^s a_{ij} \det A[\alpha^c].$$

Expanding the determinant about the ith row,

$$\text{(2.2)}\quad \det A[\alpha^c \cup \{i\} | \alpha^c \cup \{j\}] = (-1)^q a_{ij} \det A[\alpha^c] + \sum \pm a_{ip_1} a_{p_1 p_2} \cdots a_{p_t j} \det A[\alpha^c - \{p_1, \cdots, p_t\}]$$

where the summation is over all simple paths from i to j through nodes $p_1, p_2, \cdots, p_t \in \alpha^c$, $t \geqq 1$; $q = r_i + c_j$ where $r_i \langle c_j \rangle$ is now the position of row i $\langle$column $j\rangle$ of A in $A[\alpha^c \cup \{i\} | \alpha^c \cup \{j\}]$ and the ± sign in the summation depends on i, j, α and the length of the path.

It can be shown that $s + q = 2(i + j + 1)$, which is even. This can be seen by first taking $\beta = \{1, \cdots, l\} \supseteq \alpha$, and deleting one index at a time until α is obtained.

Thus, if (2.1) holds, each term in the summation in (2.2) must be zero. So either there is no path in D from i to j through α^c (condition (i)), or if such a path exists, then the complementary minor must be zero generically (condition (ii)).

Conversely, if condition (i) is true, then there is no nonzero term $a_{ip_1} a_{p_1 p_2} \cdots a_{p_t j}$ in the expansion (2.2), and so $(A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$. Alternatively if condition (ii) is true, then whenever $a_{ip_1} a_{p_1 p_2} \cdots a_{p_t j}$ is nonzero, the complementary determinant

$$\det A[\alpha^c - \{p_1, \cdots, p_t\}] = 0,$$

so in this case also $(A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$. □

Note that if $A[\alpha^c \cup \{i\} | \alpha^c \cup \{j\}]$ is reducible with respect to $A[\alpha^c]$, then (2.1) holds, but the converse is not necessarily true. It is possible that $(A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$ for noncombinatorial reasons (i.e., this equality is not generic); see Example 4.1. The fact that the result of Theorem 2.3 holds for a whole class of matrices consistent with a given D is illustrated in Example 4.6.

If D contains a self loop for each node in α^c, then the determinant in (ii) is never generically zero, and so the characterization rests solely on (i).

COROLLARY 2.4. *Assuming* (H), *given α and particular $i, j \in \alpha$, and assuming D contains a self loop on each node in α^c, then* (2.1) *holds iff j is not reachable from i through α^c.* □

This yields the following monotonicity result.

COROLLARY 2.5. *Assuming* (H), *given α and particular $i, j \in \alpha$ and β such that $\alpha \subseteq \beta \subseteq N$ with $A[\beta^c]$ nonsingular, and assuming D contains a self loop on each node in α^c, then $(A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$ (generically) implies $(A^{-1}[\beta])^{-1}_{ij} = a_{ij}$ (generically).*

Proof. Since there is no path from i to j through α^c, there is none through β^c. □

There is a vast literature concerning sparse matrix computation using graph-theoretic techniques to analyze fill-in during Gaussian elimination. Corollary 2.4 (for the case that $a_{ij} = 0$) is essentially the fundamental theoretical result upon which this sparse matrix analysis is based. Thus our main results in Theorems 2.1 and 2.3 may be viewed as part of the theoretical foundation of this analysis. The application of Corollary 2.4 to the modeling of Gaussian elimination using reachable sets may be found in [8], and indeed may be traced back to [12], but our more general theorems do not seem to be in the literature. Whereas the literature on Gaussian elimination for sparse matrices focuses on the preservation of zero entries in the Schur complement (and indeed in the LU factorization of A), our results characterize the preservation of both zero and nonzero entries.

Question (QI)(b) can now be answered by requiring the conditions of Theorem 2.3 to hold for all $i, j \in$ given α, giving necessary and sufficient conditions for the entire Schur complement of $A[\alpha^c]$ in A to be generically equal to $A[\alpha]$.

COROLLARY 2.6. *Assuming* (H) *and given α, then* (1.1) *holds* (*generically*) *iff for each $i, j \in \alpha$ either* (i) *or* (ii) *of Theorem* 2.3 *holds.* □

From Corollary 2.6, we have the following special case.

COROLLARY 2.7. *Assuming* (H), *given* α *and assuming* D *contains a self loop on each node in* α^c, *then* (1.1) *holds* (*generically*) *iff there exists a permutation matrix* P *such that*

$$P^TAP = \left[\begin{array}{c|c|c} A[\gamma] & A_{12} & A_{13} \\ \hline 0 & A[\alpha] & A_{23} \\ \hline 0 & 0 & A[\beta] \end{array}\right] \tag{2.3}$$

where $\alpha \cup \beta \cup \gamma = N$ *and either one of* β, γ *may be empty.*

Proof. This form is obtained by noting that under the hypotheses (1.1) holds (generically) iff there is no path in D from any node in α to any node in α through α^c. It can be shown that this path condition is satisfied iff N can be written as the disjoint union of α and sets β, γ such that there is no edge in D from β to α, from β to γ and from α to γ. This in turn is equivalent to the existence of a permutation matrix P such that (2.3) holds. □

Next suppose we are given particular $i, j \in N$ and want this entry to be inherited for all α containing $\{i, j\}$ (i.e., question (QI)(c)). Since we must now assume that $A[\alpha^c]$ is nonsingular for all such α, the determinant in condition (ii) of Theorem 2.3 can never be zero, so we have the following characterization.

COROLLARY 2.8. *Assuming* (H) *for all* α *containing a given* i, j, *then* (2.1) *holds for all such* α *iff there exists a permutation matrix* P *such that*

$$P^TAP = \left[\begin{array}{c|c} A_{11} & \begin{matrix} & 0 \\ a_{ij} & \end{matrix} \\ \hline A_{21} & A_{22} \end{array}\right] \tag{2.4}$$

where A_{11}, A_{21}, A_{22} *are arbitrary matrices consistent with* (H), *and* a_{ij} *is the only* (*possibly*) *nonzero entry in its off-diagonal block.*

Proof. Condition (i) of Theorem 2.3 holds for all α containing $\{i, j\}$ iff there is no path from i to j through any intermediate nodes. Equivalently, removal of the edge (i, j) from D makes A reducible and this occurs iff (2.4) holds for some permutation matrix P. □

If, in addition to the hypotheses of Corollary 2.8, A is assumed to be combinatorially symmetric, then there can be no path from j to i through any intermediate nodes; removal of the edge (i, j) in the undirected graph causes i and j to be in different connected components. (We note that such an edge is often called a bridge.) Thus there exists a permutation matrix P such that

$$P^TAP = \left[\begin{array}{c|c} A_{11} & \begin{matrix} & 0 \\ a_{ij} & \end{matrix} \\ \hline \begin{matrix} & a_{ji} \\ 0 & \end{matrix} & A_{22} \end{array}\right] \tag{2.5}$$

where A_{11} and A_{22} are arbitrary combinatorially symmetric matrices consistent with (H), and a_{ij}, a_{ji} are the only (possibly) nonzero entries in the off-diagonal blocks. Note that tridiagonal matrices with nonvanishing principal minors are of this type; and the property of tridiagonal matrices that (2.1) holds for all $\alpha \subseteq N$, $i \in \alpha$, $j = i + 1 \in \alpha$ extends to general matrices of this type. Note that in this case when A is combinatorially symmetric and $i = j$, then node i can be connected to no other node.

Our fourth submatrix question, (QI)(d), requires equality (generically) for the Schur complement of $A[\alpha^c]$ in A for every α.

COROLLARY 2.9. *Assuming* (H) *for all* α, *then* (1.1) *holds* (*generically*) *for all* α *iff there exists a permutation matrix* P *such that* P^TAP *has the form* (1.2). □

Equality in this case is equivalent to having no simple path of length $\geqq 2$ in D. If in addition A is combinatorially symmetric, then (1.1) holds (generically) for all α iff A is a diagonal matrix.

Clearly we can also use Theorem 2.3 to characterize generic equality of a particular set of submatrices. Upper triangular matrices (considered in the introduction) are an example where it is natural to consider α as any set of consecutive indices. As another example, there is a simple but useful sufficient condition for

$$(2.6) \qquad (A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$$

to hold for an entire row or column of $A[\alpha]$.

COROLLARY 2.10. *Let* $\alpha \subseteq N$ *and let* A *be a nonsingular matrix with* $A[\alpha^c]$ *nonsingular. If for some fixed* $i \in \alpha$, $a_{ik} = 0$ *for all* $k \in \alpha^c$, *then* (2.6) *holds for all* $j \in \alpha$. (*A similar result holds for a fixed* $j \in \alpha$.)

Proof. This follows from the if part of Theorem 2.3 applied to $D(A)$, since there is no edge from i to any node in α^c. □

Corollary 2.10 and its analogue show that a simple sufficient condition for (2.6) is that either $A[\{i\}|\alpha^c]$ or $A[\alpha^c|\{j\}]$ be a zero matrix. This is far from necessary in general. But in the case that $\alpha^c = \{k\}$, then the result of Corollary 2.2 reduces to $(A^{-1}[N-\{k\}])^{-1}_{ij} = a_{ij} - a_{ik}a_{kj}/a_{kk}$, and thus the vanishing of a_{ik} or of a_{kj} is necessary and sufficient for (2.6) to hold.

3. Graph containment results.

We now consider the graph containment questions raised in (QII). These focus on the zero-nonzero pattern in A, and require that the zero entries are inherited, that is, no new edge is created. Given a nonsingular matrix A and a set $\alpha \subseteq N$ with $A[\alpha^c]$ nonsingular, we write $D(A^{-1}[\alpha])^{-1} \subseteq D(A[\alpha])$ iff whenever $a_{ij} = 0$ for $i, j \in \alpha$, then $(A^{-1}[\alpha])^{-1}_{ij} = 0$. We write $D(A^{-1}[\alpha])^{-1} \subseteq D(A[\alpha])$ (generically) if for any nonsingular matrix B with $B[\alpha^c]$ nonsingular and $D(B) = D(A)$ we have $D(B^{-1}[\alpha])^{-1} \subseteq D(B[\alpha])$. Graph containment for the undirected graph of a combinatorially symmetric matrix is defined similarly. On restricting $D = D(A)$ in Theorem 2.3, the following result answers question (QII)(a) and thereby specifies conditions for which the digraph of the Schur complement $(A^{-1}[\alpha])^{-1}$ is a subgraph of the digraph of $A[\alpha]$.

COROLLARY 3.1. *Given nonsingular* A *and* $\alpha \subseteq N$ *with* $A[\alpha^c]$ *nonsingular, then* $D(A^{-1}[\alpha])^{-1} \subseteq D(A[\alpha])$ (*generically*) *iff for each* $i, j \in \alpha$ *such that* $a_{ij} = 0$ *either* (i) *or* (ii) *of Theorem* 2.3 *holds with respect to* $D = D(A)$. □

In case $\alpha^c = \{k\}$, we have digraph containment iff there is an edge (i, j) in $D(A)$ whenever there are edges (i, k) and (k, j), and a self loop at node i whenever there are edges (i, k) and (k, i). In the terminology of [13], this containment condition is equivalent to the *deficiency* of k equal to the empty set and node k not in any 2-cycle $i \to k \to i$ with $a_{ii} = 0$.

On restricting A to be combinatorially symmetric with all diagonal entries nonzero, condition (ii) is not required in Corollary 3.1 (as we are considering the "generic" inheritance of zeros); this provides an answer to (QII)(c). In the case that α^c is a single node k, we obtain an answer to this question which involves a well-known concept from the study of Gaussian elimination on sparse matrices (see, e.g., [8]). The remark following Corollary 2.10 allows us to omit "generically" here.

COROLLARY 3.2. *Let A be an n-by-n nonsingular combinatorially symmetric matrix having all $a_{kk} \neq 0$. Then for $k \in N$*

$$G(A^{-1}[N-\{k\}])^{-1} \subseteq G(A[N-\{k\}])$$

iff every two neighboring nodes of k are connected by an edge in $G(A)$. □

Equivalently, the above condition can be stated as node k is *simplicial* in $G(A)$ (see, e.g., [10]), and Corollary 3.2 embodies the well-known fact that pivoting on a simplicial node causes no fill-in in Gaussian elimination (that is, no zero entries become nonzero). However, when $|\alpha^c| > 1$, it is possible that none of the nodes in α^c is simplicial (see Example 4.4), but that the undirected graph containment holds. In order to obtain a necessary and sufficient condition in this case, we introduce a new definition. A set of connected nodes V of an undirected graph G is called *simplicial in G* if the set of all nodes not in V and adjacent to any node in V induces a complete subgraph in G.

THEOREM 3.3. *Let A be an n-by-n nonsingular combinatorially symmetric matrix having all $a_{kk} \neq 0$. For a given $\alpha \subseteq N$ such that $A[\alpha^c]$ is nonsingular, let $G_{\alpha^c}(A)$ denote the subgraph of $G(A)$ induced by the node set α^c. Suppose $\alpha^c = \bigcup_{k=1}^{m} \beta_k$ where β_k are mutually disjoint and the subgraphs $G_{\beta_k}(A)$ are the connected components of $G_{\alpha^c}(A)$. Then*

$$G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha]) \qquad (generically)$$

iff each set of nodes β_k $(1 \leqq k \leqq m)$ is simplicial in $G(A)$.

Proof. Suppose first that each set of nodes β_k is simplicial in $G(A)$. If $i, j \in \alpha$ with $a_{ij} = 0$, then this implies that node j is not connected to node i by a path with nodes solely in α^c. Thus $G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha])$, by Theorem 2.3.

Conversely, if $G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha])$ (generically), then (as all $a_{kk} \neq 0$) condition (i) of Theorem 2.3 must hold for all $i, j \in \alpha$ such that $a_{ij} = 0$. Thus any two nodes in α which are adjacent to nodes in some set β_k must be connected by an edge, i.e., each set β_k is simplicial in $G(A)$. □

Corollary 3.1 does not seem to have been stated in the literature, although the graph structure of the Schur complement is considered in [2]. We note that this graph structure is important in partitioned (or block) methods of solving sparse linear systems (see [2], [5]). Corollary 3.2 is also well known (see, e.g., [8]), but our generalization (Theorem 3.3) and the concept of simplicial sets of nodes is new.

Coming now to our final pair of questions (QII)(b),(d) we have to consider all index sets α. In the directed graph case, if digraph containment holds for every choice of $\alpha \subseteq N$, then $A[\alpha^c]$ must be nonsingular for all α, implying that all $a_{kk} \neq 0$. Thus $D(A^{-1}[\alpha])^{-1} \subseteq D(A[\alpha])$ (generically) iff the deficiency of each node is empty. Note that the deficiency of each node being empty means that the graph $D(A)$ is transitively closed. In this event $D(A^{-1}) \subseteq D(A)$ (see, e.g., [9]). If $D(A)$ is transitively closed, then so is $D(A[\alpha])$, and thus $D(A^{-1}[\alpha])^{-1} \subseteq$ the transitive closure of $D(A^{-1}[\alpha]) \subseteq D(A[\alpha])$. Similar reasoning via transitive closure verifies the converse, providing an alternate elementary verification of the answer to (QII)(b).

For the undirected case we have the following theorem.

THEOREM 3.4. *Let A be an n-by-n nonsingular combinatorially symmetric matrix having all principal minors nonvanishing. Then*

$$G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha]) \qquad for\ every\ \alpha \subseteq N$$

iff each node is simplicial, that is iff $G(A)$ is a direct sum of complete graphs.

Proof. If the graph containment holds, then using Corollary 3.2 with $\alpha^c = \{k\}$ each node k must be simplicial. Conversely, if each node is simplicial, every connected set of nodes in $G(A)$ must form a complete subgraph. Therefore the subgraph induced by the neighbors of any connected subset V of nodes of $G(A)$ must also form a complete subgraph, and thus V is a simplicial set of nodes in $G(A)$. Theorem 3.3 gives the graph containment. □

Finally, we mention a different approach to inherited zeros, one which utilizes the structure of A^{-1} corresponding to a given zero pattern in A. The main tool is the following rank result recently proved independently (see [6, Cor. 3]).

THEOREM 3.5. *Let A be a nonsingular n-by-n matrix and let $\beta, \gamma \subseteq N$ with cardinalities p and q, respectively. Then*

$$\operatorname{rank} A^{-1}[\gamma^c|\beta^c] = \operatorname{rank} A[\beta|\gamma] + n - p - q. \qquad \Box$$

Application of this identity gives a less direct method of ascertaining inherited zeros than the results above, but in some cases yields insights that the graph theoretic approach does not. It can also be used to give simpler and more informative proofs for Theorems 2.1 and 3.1 in [1]. We illustrate the use of Theorem 3.5 in Example 4.7.

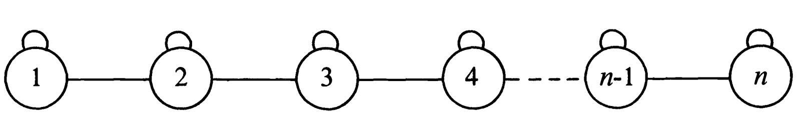
4. Examples. We now give examples to illustrate our results and answers to questions (QI) and (QII).

Example 4.1. We note that if $(A^{-1}[\alpha])^{-1}_{ij} = a_{ij}$ (in which the equality is not necessarily generic), then neither (i) nor (ii) of Theorem 2.3 necessarily holds, as the following example shows. Let $\alpha = \{3, 4\}$ and

$$A = \begin{bmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 \\ 2 & 6 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \text{so } A^{-1} = \begin{bmatrix} 1 & -5 & 1 & 3 \\ 0 & 1 & 0 & -1 \\ -2 & 4 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$$

and $(A^{-1}[\alpha])^{-1}_{34} = a_{34} = 0$. The preservation of this zero entry is due to the fact that $\det A[\{1, 2, 3\}|\{1, 2, 4\}] = 0$ because of the numerical values of the entries, *not* because of the graph $D = D(A)$. Relative to D this is an example of "chance cancellation," rather than a generic identity.

Example 4.2. Let G denote a "straight-chain" graph on n nodes with a self loop at each node:



Let A be any n-by-n nonsingular combinatorially symmetric matrix which has all $a_{kk} \neq 0$ and $G(A) = G$. Let $\alpha = \{p, p + 1, \cdots, q\}$ where $1 < p \leqq q < n$.

If $A[\alpha^c]$ is nonsingular, then using Theorem 2.3 with the corresponding $D(A)$ implies that (1.1) holds except for a_{pp} and a_{qq} (as the only paths through α^c between two nodes in α are cycles from node p to node p, and from node q to node q).

By Theorem 3.3, $G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha])$ (generically) since $G_{\alpha^c}(A)$ has two connected components (one with node set $\{1, 2, \cdots, p - 1\}$ and the other with node set $\{q + 1, q + 2, \cdots, n\}$) and both of these node sets are simplicial sets in $G(A)$.

Example 4.3. Let G and A be as in Example 4.2, but now consider

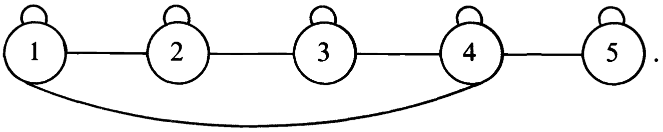
$$\alpha^c = \{p, p + 1, \cdots, q\},$$

where $1 < p \leqq q < n$. Suppose $A[\alpha^c]$ is nonsingular. By Theorem 2.3, equality (1.1) holds except for the entries $a_{p-1,p-1}$, $a_{p-1,q+1}$, $a_{q+1,p-1}$ and $a_{q+1,q+1}$. As

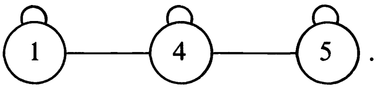
$$a_{q+1,p-1} = a_{p-1,q+1} = 0$$

and these entries become nonzero in $(A^{-1}[\alpha])^{-1}$, clearly $G(A^{-1}[\alpha])^{-1} \not\subseteq G(A[\alpha])$ (generically). This can also be seen from Theorem 3.3 since the set of nodes $\{p, p+1, \cdots, q\}$ is not a simplicial set (as its adjacent nodes $p-1$ and $q+1$ are not connected). As noted in the Introduction, $(A^{-1}[\alpha])^{-1}$ is also tridiagonal; thus, the zero pattern is preserved despite the lack of graph containment.

Example 4.4. Let A be a nonsingular combinatorially symmetric matrix with the following undirected graph:



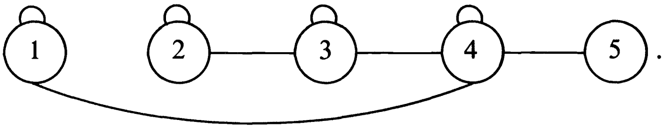
Let $\alpha = \{1, 4, 5\}$. If $A[\alpha^c]$ is nonsingular, then $G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha])$ (generically) by Theorem 3.3 as the set of nodes $\{2, 3\}$ is a simplicial set in $G(A)$. Specifically, the zero entries a_{15} and a_{51} are inherited by $(A^{-1}[\alpha])^{-1}$, and $G(A^{-1}[\alpha])^{-1}$ is as follows:



Note that neither node 2 nor 3 is simplicial in $G(A)$.

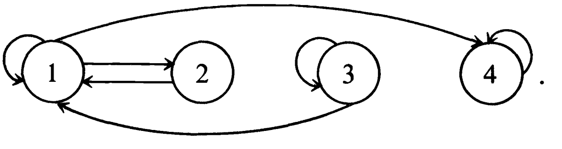
Example 4.5. In this example only, we consider the undirected graph of a matrix that has a zero entry on the diagonal. With respect to the result of Theorem 3.3, each set of vertices β_k simplicial in $G(A)$ implies $G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha])$ (generically) even when some $a_{kk} = 0$ for $k \in \alpha^c$. However, the converse of this result does not follow, as this example illustrates.

Let A be a nonsingular combinatorially symmetric matrix with the following undirected graph:



Letting $\alpha = \{1, 2\}$, the set of nodes $\alpha^c = \{3, 4, 5\}$ is not a simplicial set, however, $G(A^{-1}[\alpha])^{-1} \subseteq G(A[\alpha])$ (generically) as the zero entries are inherited, using the fact that $a_{55} = 0$ (cf. with Example 4.4).

Example 4.6. Consider the following directed graph D:



Let $\alpha = \{3, 4\}$. If A is a nonsingular matrix with $D(A) = D$ and $A[\alpha^c]$ is nonsingular, then either Theorem 2.3 or Corollary 2.6 implies that $(A^{-1}[\alpha])^{-1} = A[\alpha]$ (generically) since the only path between two nodes in α passing through α^c is $3 \rightarrow 1 \rightarrow 4$; however, $\det A[\alpha^c - \{1\}] \equiv a_{22} = 0$. This remains true for all A consistent with D such that $A[\alpha^c]$ is nonsingular, specifically any or all of a_{11}, a_{14}, a_{31} can be set to zero.

Thus, the Schur complement of $A[\{1, 2\}]$ in A is identically equal to the diagonal submatrix $A[\{3, 4\}]$. It is interesting to note, however, that the (3, 4) entry of the Schur complement of $A[\{1\}]$ in A is nonzero (as can be seen from the remark following Corollary 2.10).

Example 4.7. Let A be a 5-by-5 nonsingular matrix with $a_{13} = a_{15} = a_{43} = a_{45} = 0$ and remaining entries arbitrary. Let $\alpha = \{1, 2, 3\}$, $\beta = \{1, 4\}$ and $\gamma = \{3, 5\}$ and assume $A[\alpha^c]$ is nonsingular. Clearly rank $A[\beta|\gamma] = 0$, so rank $A^{-1}[\gamma^c|\beta^c] = 1$ by Theorem 3.5. Thus rank $(A^{-1}[\alpha])[\{1, 2\}|\{2, 3\}] \leqq 1$ implying that rank $(A^{-1}[\alpha])^{-1}[\{1\}|\{3\}] \leqq 0$ by Theorem 3.5; thus the zero entry a_{13} is inherited. This also follows from Theorem 2.3(i) with $D = D(A)$.

REFERENCES

[1] W. W. Barrett and P. J. Feinsilver, *Inverses of banded matrices*, Linear Algebra Appl., 41 (1981), pp. 111–130.

[2] J. R. Bunch, *Block methods for solving sparse linear systems*, in Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 39–58.

[3] D. Carlson, *What are Schur complements, anyway*? Linear Algebra Appl., 74 (1986), pp. 257–275.

[4] D. Carlson and T. L. Markham, *Schur complements of diagonally dominant matrices*, Czechoslovak Math. J., 29 (1979), pp. 246–251.

[5] I. S. Duff, *Research directions in sparse matrix computations*, in MAA Studies in Math., 24, Studies in Numerical Analysis, Gene H. Golub, ed., 1984, pp. 83–139.

[6] M. Fiedler and T. L. Markham, *Completing a matrix when certain entries of its inverse are specified*, Linear Algebra Appl., 74 (1986), pp. 225–237.

[7] F. R. Gantmacher, *Matrix Theory, Vol.* I, Chelsea, New York, 1959.

[8] J. A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[9] J. R. Gilbert, *Predicting structure in sparse matrix computations*, Cornell Univ. Report CS-86-750, Cornell Univ., Ithaca, NY, 1986.

[10] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[11] R. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, London, 1985.

[12] S. Parter, *The use of linear graphs in Gauss elimination*, SIAM Rev., 3 (1961), pp. 119–130.

[13] D. J. Rose and R. E. Tarjan, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM J. Appl. Math., 34 (1978), pp. 176–197.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

DIGRAPH DECOMPOSITIONS AND EULERIAN SYSTEMS*

ANDRÉ BOUCHET†

Abstract. The theory of digraph decompositions introduced by W. Cunningham and the theory of isotropic systems introduced by the author are unified. A basic combinatorial tool is the operation of local complementation at a vertex of a digraph, a generalization of the similar operation already known for simple graphs. This allows us to unify in a single class the semibrittle digraphs characterized by W. Cunningham and to devise a more efficient algorithm for searching for a split of a digraph.

Key words. digraph decomposition, isotropic system, Eulerian system

AMS(MOS) subject classification. 05C

1. Introduction. Graphs and digraphs considered throughout this paper will be finite and simple. Thus a *digraph* (*graph*) is defined by a finite vertex-set $V(G)$ and an arc-set $A(G) \subseteq \{(x, y) : x, y \in V(G), x \neq y\}$ (edge-set $E(G) \subseteq \{\{x, y\} : x, y \in V(G), x \neq y\}$). As usual the notation $\{x, y\}$ for an edge will be simplified into xy. It will be convenient to consider graphs as special cases of digraphs by identifying any edge xy with the pair of reversed arcs $\{(x, y), (y, x)\}$. If v is a vertex of a digraph G then the subgraph of G induced on $V(G)\backslash\{v\}$ will be denoted by $G\backslash v$. For the most part, our terminology and notation follows [1].

A *simple decomposition* of a digraph G is a pair $\{G', G''\}$ of digraphs satisfying the following properties: (i) $|V(G')|, |V(G'')| \geqq 3$; (ii) $V(G') \cap V(G'')$ contains precisely one vertex v called the *marker*; (iii) $V(G) = V(G'\backslash v) \cup V(G''\backslash v)$; (iv) $D(G) = D(G'\backslash v) \cup D(G''\backslash v) \cup \{(x, y) : (x, v) \in D(G') \text{ and } (v, y) \in D(G'') \text{ or } (x, v) \in D(G'') \text{ and } (v, y) \in D(G')\}$. Condition (i) is introduced essentially to avoid trivialities. Nonsimple decompositions can be defined inductively from simple ones, but we shall not consider them in this paper, and the term decomposition will always mean simple decomposition.

Decompositions of digraphs have been introduced in [8] by W. Cunningham. They encompass the notion of graph separability and the substitution decomposition or X-join. Moreover, there are a number of combinatorial optimization problems which can be solved more efficiently on each member of a decomposition. This is the case, for example, for the optimal stable set problem for simple graphs. The reader is referred to the Introduction of Cunningham's paper for details. We want to discuss another application to the recognition of circle graphs.

A *circle graph* is the intersection graph of a finite number of chords of a circle. Three independent efficient algorithms have recently been found for recognizing circle graphs [4], [5], [10], [11]. In each case the property that a simple graph G with a decomposition $\{G', G''\}$ is a circle graph if and only if G' and G'' are themselves circle graphs is used in a first step. Thus the problem is reduced to the indecomposable graphs, also called *prime graphs*, for which it can be proved they are uniquely realizable by chords of a circle [4], [5], [10]. This unique realization property and some way for reducing prime graphs are then the main lines of the algorithms described in [4], [5], [10].

We have recently introduced the notion of an isotropic system [2], [3] as a means for unifying some properties of 4-regular graphs and some autodual properties of binary matroids. It appears, moreover, that each isotropic system is associated to a class of

* Received by the editors November 4, 1985; accepted for publication September 19, 1986.

† Centre de Recherche en Informatique et Combinatoire, Université du Maine, 72017 Le Mans Cedex, France.

simple graphs defined up to local complementation, where the *local complementation* of a simple graph G at one of its vertices v is the operation which consists of replacing the subgraph induced by G on $n(v) = \{w : vw \in E(G)\}$ by the complementary subgraph. We have proved in [5] that the splits (see § 6) of a simple graph are invariant under local complementation, which implies that a prime graph remains prime after a local complementation.

Our purpose in this paper is to extend to arbitrary digraphs the properties which have been derived when studying isotropic systems. For that we shall define the Eulerian systems which are a relaxation of isotropic systems and we shall extend the definition of local complementations to digraphs. So we shall prove that the semibrittle digraphs described by W. Cunningham constitute a single class of digraphs under local complementation. Moreover the use of local complementation will allow us to improve the time-complexity of Cunningham's algorithm for finding a split of a digraph. Apart from these main applications, we shall establish some properties which will be used in future papers.

2. Basic definitions and notation. If V is a finite set we consider its set of parts, $\mathbf{P}(V)$, as a vector-space over $GF(2)$ where the addition is the symmetric difference. For any $v \in V$ we shall frequently simplify the notation $\{v\}$ into v. Thus in further computations, an expression like $H + v$ where $H \subseteq V$ and $v \in V$ will be read as $H \triangle \{v\}$.

If E and V are two sets and if we have defined an operation $(e, v) \twoheadrightarrow e*v$ from $E \times V$ into E then, where V^* is the free monoid on V, we shall implicitly extend the operation from $E \times V^*$ into E by means of the recursive formula $(e*m)*v = e*mv$ for every $m \in V^*$ and $v \in V$.

Throughout this paper we consider a fixed vector-space K of dimension 2 over $GF(2)$. Thus K has precisely one null element and three nonnull elements. We let$K' = K\backslash\{0\}$. For any finite set V we shall consider K^V as a vector-space of dimension $2|V|$ over $GF(2)$. The *support* of a vector $A \in K^V$ is $\{v : A(v) \neq 0\}$. A *complete vector* is a vector with a support equal to V. In other terms a complete vector is an element of K'^V. Two *supplementary vectors* are two complete vectors A and B such that $A(v) \neq B(v)$ for every $v \in V$. For a vector $A \in K^V$ and a subset $W \subseteq V$, we denote by AW the vector of K^V defined by $AW(v) = A(v)$ if $v \in W$ and $AW(v) = 0$ if $v \notin W$. We shall denote by $\hat{A}$ the set $\{AW : W \subseteq V\}$. This is a subspace of dimension $|V|$ in K^V.

By convention, for any vector $A \in K^V$ and for any algebraic expression Exp whose value is a subset $P \subseteq V$, the value of the expression $A[\text{Exp}]$ is equal to AP—the expression $A(\text{Exp})$ would be ambiguous. For example, where W' and W'' are subsets of V, we have the equality $AW'W'' = A[W' \cap W'']$.

Let $S = (L, V)$ be a pair with a finite set V and a subspace L of K^V. A complete vector A is called an *Eulerian vector* for S if $\hat{A}$ and L are disjoint. The pair S is called an *Eulerian system* if it admits some Eulerian vector and if dim $(L) = |V|$. We notice that the condition dim $(L) = |V|$ is not sufficient for the existence of an Eulerian vector. A counterexample is constructed by choosing V and $W \subseteq V$ satisfying $|W| = |V|/2$ and letting $L = K^W$.

Isotropic systems are introduced in [2], [3]. They are defined in the following way. We consider on K the bilinear form $(x, y) \twoheadrightarrow xy$ such that $xy = 1$ if and only if $0 \neq x \neq y \neq 0$. For any finite set V we define on the vector-space K^V the bilinear form$(A, B) \twoheadrightarrow AB = \Sigma(A(v)B(v) : v \in V)$. We have clearly that $AA = 0$ for any $A \in K^V$. An *isotropic system* is a pair $S = (L, V)$ where L is a subspace of K^V which is totally isotropic, i.e., $AB = 0$ whenever $A, B \in L$— and is of dimension $|V|$. It is proved in [2] that every isotropic system has some Eulerian vector. Therefore Eulerian systems generalize isotropic systems. How to associate isotropic systems to 4-regular graphs is also described, and it appears that Eulerian vectors correspond bijectively to Eulerian tours, which explains the present terminology.

3. Graphic presentations and switching property. The *neighborhood function* of a digraph F is the linear endomorphism n of $\mathbf{P}(V)$ satisfying $n(v) = \{w : (v, w) \in D(F)\}$ for every vertex v. Thus for any $W \in \mathbf{P}(V)$ we have $n(W) = \Sigma(n(v) : v \in W)$.

PROPOSITION 3.1. *Let (F, A, B) be a triple with a digraph F on the vertex-set V and two supplementary vectors A and B of K^V, and let n be the neighborhood function of F. Then*

(i) *The mapping $\alpha : W \twoheadrightarrow An(W) + BW$ is a linear injection from* $\mathbf{P}(V)$ *into K^V (let L =* $\operatorname{Im}(\alpha)$);

(ii) *The pair $S = (L, V)$ is an Eulerian system, and A is an Eulerian vector of S.*

Proof. The mapping α is clearly linear from $\mathbf{P}(V)$ into K^V. Therefore L is a subspace of K^V. For every $W \subseteq V$ and every $v \in W$, $\alpha(W)(v)$ is either equal to $B(v)$ or $B(v) + A(v)$. Since A and B are supplementary, we have always $0 \neq \alpha(W)(v) \neq A(v)$. This implies two consequences. First, the kernel of α is reduced to the empty set, which proves (i) and implies dim $(L) = |V|$. Second, no nonnull vector of L is in $\hat{A}$, and so A is an Eulerian vector of S. □

DEFINITION. The triple (F, A, B) will be called a *graphic presentation* of S, and F a *fundamental graph* of S. The image by α of the canonical base of $\mathbf{P}(V)$ will be called the *fundamental base* of L induced by A. This fundamental base is $\{An(v) + Bv : v \in V\}$.

Two complete vectors A and B of K^V are said to be *neighbours* at $v \in V$ if they satisfy $A(v) \neq B(v)$ and $A(w) = B(w)$ for every $w \neq v$ in V. A subset $\Sigma \subseteq K'^V$ satisfies the *switching property* if for every $A \in \Sigma$ and every $v \in V$ there exists precisely one vector $B \in \Sigma$ which is neighbour of A at v. Then we shall say that B is obtained by *switching A* at v, and it will be denoted as $A*v$. Since $|K'| = 3$, the switching property means equivalently that for every $A \in \Sigma$ and every $v \in V$ there exists precisely one complete vector $C \notin \Sigma$ which is neighbour of A at v.

PROPOSITION 3.2. *The set of the Eulerian vectors of an Eulerian system satisfies the switching property.*

Proof. Let $S = (L, V)$ be the Eulerian system. Let A be an Eulerian vector of S and A' and A'' be the two complete vectors which are neighbours of A at v. We prove first that A' and A'' cannot be both Eulerian. If this was the case, L would be disjoint from $X = \hat{A} \cup \hat{A}' \cup \hat{A}''$. But X is the subspace of all the vectors $a \in K^V$ satisfying $a(w) = A(w)$ or 0 for every $w \in V\backslash v$, and $a(v)$ is arbitrary in K. This subspace has a dimension equal to $|V| + 1$, and so it cannot be disjoint of L whose dimension is equal to $|V|$. Finally we prove the impossibility that A' and A'' can both be noneulerian. If this were the case there would exist some vectors $a' \in \hat{A}' \cap L$ and $a'' \in \hat{A}'' \cap L$. We have necessarily $a'(v) = A'(v)$ since otherwise $a'(v)$ would be null and so a' would belong to $\hat{A}$, a contradiction since A is Eulerian. Similarly we have $a''(v) = A''(v)$. Let $a = a' + a''$. We have$a(v) = A'(v) + A''(v) = A(v)$, and $a(w) = A(w)$ or 0 for every $w \in V\backslash v$. Therefore $a \in \hat{A}$, a contradiction since A is Eulerian. □

PROPOSITION 3.3. *For every Eulerian vector A of an Eulerian system $S = (L, V)$ there exists precisely one graphic presentation (F, A, B).*

Proof. Following the switching property, there exists for each $v \in V$ precisely one complete noneulerian vector A'_v which is a neighbour of A at v. Let B be the complete vector defined by $B(v) = A'_v(v)$ for each $v \in V$. The subspace $\hat{A}'_v \cap L$ contains some nonnull vector A_v. This vector satisfies $A_v(w) = A(w)$ or 0 for every $w \neq v$ in V, and $A_v(v) = B(v)$ or 0. The equality $A_v(v) = 0$ is ruled out since otherwise A_v would be an element of $\hat{A}$ when A is Eulerian. If C_v is another vector of $\hat{A}'_v \cap L$ it also satisfies $C_v(v) = B(v)$. The vector $D_v = A_v + C_v$ satisfies $D_v(v) = 0$ and $D_v(w) = A(w)$ or 0. This implies that $D_v \in \hat{A}$, and $D_v = 0$ because A is Eulerian. Therefore A_v is the single nonnull vector belonging to $\hat{A}'_v \cap L$.

Let F be the digraph defined by $V(F) = V$ and $A(F) = \{(v, w) : A_v(w) = A(w)\}$. If n is the neighborhood function of F, it is easy to verify that we have $A_v = An(v) + Bv$ for every $v \in V$. This implies that $An(W) + BW \in L$ for every $W \subseteq V$. Therefore the Eulerian system $S' = (L', V)$ induced by the graphic presentation (F, A, B) satisfies $L' \subseteq L$. The equality holds because L and L' have the same dimension.

Let (F', A, B') be another graphic presentation of S, and let n' be the neighborhood function of F'. For every $v \in V$ let A''_v be the complete vector which is a neighbour of A at v and satisfies $A''_v(v) = B'(v)$. The vector $An'(v) + B'v$ belongs to L because it is induced by the graphic presentation (F', A, B'). This vector is contained in $\hat{A}''_v$. Therefore A''_v is noneulerian. Since there is a single noneulerian vector which is a neighbour of A at v, we have $A'_v = A''_v$. This implies that $B' = B$. Finally we have already noticed that for every $v \in V$, there is a single nonnull vector in $\hat{A}'_v \cap L$. This implies that $n'(v) = n(v)$, and thus $F' = F$. □

We shall say that the graphic presentation (F, A, B) is induced by the Eulerian vector A.

The results proved in [3] imply that an Eulerian system S is an isotropic system if and only if there exists some fundamental digraph of S which is a simple graph. In fact every fundamental digraph of S will be in this case a simple graph. We notice also that (3.2) and (3.3) are direct generalizations of similar properties of isotropic systems.

4. Local complementation. If $GP = (F, A, B)$ is a graphic presentation of an Eulerian system, and $v \in V$, we denote by $GP * v$ the graphic presentation induced by $A * v$ and we shall say that $GP * v$ is obtained by switching GP at v. In this section we give the formulas for computing $GP * v$.

Let F be a digraph on the vertex-set V. A pair (x, y) of distinct vertices that is not an arc of F will be called a *coarc*. A *transitivity arc* (*transitivity coarc*) at a vertex v is an arc (a coarc) (x, y) such that $v \neq x \neq y \neq v$, (x, v) and (v, y) are arcs of F. To *locally complement* F at v is to exchange the transitivity arcs at v with the transitivity coarcs at v. The resulting digraph will be denoted as $F * v$. It is easy to verify that the present definition of local complementation encompasses those for simple graphs. The following property is the generalization to Eulerian systems of a similar one proved in [3] for isotropic systems.

PROPOSITION 4.1. *Let* $GP = (F, A, B)$ *be a graphic presentation of an Eulerian system* $S = (L, V)$ *and let* $v \in V$. *If* n *is the neighborhood function of* F *then* $GP * v = (F * v, A + Bv, B + A[n(v) \cap n^{-1}(v)])$.

Proof. Let $\{A_u = An(u) + Bu : u \in V\}$ be the fundamental base of L induced by A. We define a new base $\{A'_u : u \in V\}$ by a kind of pivoting at v. We let

$$A'_v = A_v,$$

$$A'_u = A_u + A_v \quad \text{if } u \neq v \text{ and } A_u(v) \neq 0,$$

$$A'_u = A_u \quad \text{if } u \neq v \text{ and } A_u(v) = 0.$$

Expressed in terms of the neighborhood function n, this yields

$$A'_v = An(v) + Bv,$$

$$A'_u = An(u) + An(v) + Bu + Bv \quad \text{if } u \neq v \text{ and } v \in n(u),$$

$$A'_u = An(u) + Bu \quad \text{if } u \neq v \text{ and } v \notin n(u).$$

Let $n'' : V \twoheadrightarrow \mathbf{P}(V)$ be defined by

$$n''(v) = n(v),$$
$$n''(u) = n(u) + n(v) \quad \text{if } u \neq v \text{ and } v \in n(u),$$
$$n''(u) = n(u) \quad \text{if } u \neq v \text{ and } v \notin n(u).$$

If we let $A' = A + Bv$, then the reader will easily verify that

$$A'_u = A'n''(u) + Bu \quad \text{for every } u \in V.$$

It is also easy to verify that the neighborhood function n' of $F' = F * v$ satisfies

$$n'(u) = n''(u) + u \quad \text{if } u \in n(v) \cap n^{-1}(v),$$
$$n'(u) = n''(u) \quad \text{otherwise.}$$

Thus if we let $B' = B + A[n(v) \cap n^{-1}(v)]$, we find that

$$A'_u = A'n'(u) + B'u \quad \text{for every } u \in V.$$

Therefore if $S' = (L', V)$ is the Eulerian system induced by the graphic presentation (F', A', B'), then $\{A'_u : u \in V\}$ is the fundamental base of L' induced by A', which implies $L' = L$. □

Remark. The switching $A * v$ of a complete vector A at a vertex v is defined with respect to a given Eulerian subset $\Sigma \subseteq K'^V$ satisfying the switching property. On the other hand, the local complementation $F * v$ of a digraph at one of its vertices v is defined absolutely. Similarly the switching $GP * v$ of graphic presentation GP is defined absolutely. The preceding proposition tells us that GP and $GP * v$ are graphic presentations of a same Eulerian system. We prove in the next section that any graphic presentation of an Eulerian system is accessible from a given one by a succession of switchings.

5. Switching property and accessibility. For every finite set V and any two vectors A and B of K^V we consider the Hamming distance $d(A, B) = |\{v : A(v) \neq B(v)\}|$. The length of a word $m \in V^*$ is denoted as $|m|$.

PROPOSITION 5.1. Accessibility property. *If V is a finite set, and $\Sigma \subseteq K'^V$ satisfies the switching property, then for every pair of distinct vectors A and B of Σ there exists $m \in V^*$ such that $B = A * m$ and $|m| \leqq 2d(A, B) - 1$.*

Proof. We proceed by induction on $d = d(A, B)$. If $d = 1$ the result is an immediate consequence of the switching property. Let us consider two vectors A and B in K'^V such that $d(A, B) = d > 1$. Let $K' = \{x, y, z\}$. Let $v \in V$ be such that $A(v) \neq B(v)$. We suppose that $A(v) = x$ and $B(v) = y$. Let $A' = A * v$ and $B' = B * v$. The value of $A'(v)$ is either equal to y or z. In the first case $d(A', B) = d - 1$, and by induction there exists $m' \in V^*$ such that $B = A' * m'$ and $|m'| \leqq 2d - 3$. Thus $B = A * vm'$, and the proposition is proved with $m = vm'$. The value of $B'(v)$ is either equal to x or z. In the first case we proceed as before. Thus it remains the case where $A'(v) = B'(v) = z$. We have then $d(A', B') = d - 1$. By induction there exists $m' \in V^*$ such that $B' = A' * m'$ and $|m'| \leqq 2d - 3$. Thus $B = A * vm'v$, and the proposition holds with $m = vm'v$. □

COROLLARY 5.2. *For any two Eulerian vectors A and B (graphic presentations GP and GQ, fundamental graphs F and G) of an Eulerian system $S = (L, V)$ there exists $m \in V^*$ such that $B = A * m$ ($GQ = GP * m$, $G = F * m$) and $|m| \leqq 2|V| - 1$.* □

DEFINITION. Two digraphs F and G on the same vertex-set V are *locally equivalent* if there exists $m \in V^*$ such that $G = F * m$. The preceding corollary says that the fundamental digraphs of an Eulerian system constitute a class of local equivalence. Since any digraph is a fundamental digraph of some Eulerian system we have the following result in graph theory.

COROLLARY 5.3. *For any two locally equivalent digraphs F and G on a same vertex-set V there exists $m \in V^*$ such that $G = F*m$ and $|m| \leqq 2|V| - 1$.* □

PROPOSITION 5.4. *For every Eulerian system S there exists an integer $k > 0$ such that any fundamental digraph F of S appears in precisely k graphic presentations of S.*

Proof. Let $S = (L, V)$. We consider the simple graph Q defined by $V(Q) = \{g : g$ is a graphic presentation of $S\}$, $E(Q) = \{gh : g, h \in V(Q)$ and there exists $v \in V$ such that $h = g*v\}$. We consider also the simple graph Q' defined by $V(Q') = \{g' : g'$ is a fundamental graph of $S\}$, $E(Q') = \{g'h' : g', h' \in V(Q')$ and there exists $v \in V$ such that $h' = g'*v\}$. The mapping $g \twoheadrightarrow g'$ from $V(Q)$ onto $V(Q')$, where g' is the fundamental graph of S occurring in g, is a covering mapping of Q over Q'. Following (5.2) the graph Q is connected. Therefore the statement holds with the index k of the covering mapping. □

DEFINITION. The integer k in the preceding proposition will be called the *index* of the Eulerian system S.

6. Cut-matrices. For an Eulerian system $S = (L, V)$ and a subset $V' \subseteq V$ we denote by $L \times V'$ the subspace constituted by the vectors of L whose support is contained in V', and we set $c(V') = |V'| - \dim (L \times V')$.

For a digraph F on the vertex-set V and a subset $V' \subseteq V$, the *cut-matrix* of V' is the binary matrix $\pi = (\pi_{v'v''})_{v' \in V', v'' \in V \setminus V'}$ such that $\pi_{v'v''} = 1$ if and only if (v', v'') is an arc of F.

PROPOSITION 6.1. *Let F be a fundamental digraph of an Eulerian system $S = (L, V)$, and let π be the cut-matrix in F of a subset $V' \subseteq V$. Then $c(V') = \operatorname{rank}(\pi)$.*

Proof. The proof is the same as the proof of (3.1) in [6]. □

COROLLARY 6.2. $c(V') \geqq 0$. □

COROLLARY 6.3. *Local complementations do not change the ranks of cut-matrices.* □

If V' is a subset of vertices of the digraph F, we set

$$\delta(V') = \{(v', v'') \in A(F) : v' \in V', v'' \in V \setminus V'\}.$$

We note that $\delta(V') = \varnothing$ if and only if the cut-matrix of V' has a null rank.

The *transitive closure* of F is the digraph F^t defined by $V(F^t) = V(F)$, $A(F^t) = \{(v, w) : v \neq w$, there exists a directed path from v to w in $F\}$.

COROLLARY 6.4. *Local complementations do not change the transitive closure of a digraph.*

Proof. There exists a path from a vertex v' to a vertex v'' of a digraph F if and only if for any $V' \subseteq V(F)$ which contains v' and not v'', we have $\delta(V') \neq \varnothing$. This means that the cut-matrix of V' has a nonnull rank, and this property does not change after a local complementation. □

A digraph F is said to be *diconnected* if its transitive closure is a complete digraph. This means that for every proper subset V' of vertices, $\delta(V')$ is nonempty, or equivalently, the cut-matrix of V' has a nonnull rank. We shall say that the Eulerian system S is diconnected if $c(V') > 0$ for every proper subset V' of V. From now on we shall deal only with diconnected digraphs and diconnected Eulerian systems.

A *split* of the digraph F is a bipartition $\{W', W''\}$ of its vertex-set such that $|W'|$, $|W''| \geqq 2$ and there exists two subsets W'^{+} and W'^{-} of W' and two subsets W''^{+} and W''^{-} of W'' satisfying $\delta(W') = W'^{+} \times W''^{-}$ and $\delta(W'') = W''^{+} \times W'^{-}$.

PROPOSITION 6.5. *A bipartition* $\{W', W''\}$ *of the vertex-set of a diconnected digraph* F *is a split if and only if* $|W'|, |W''| \geqq 2$ *and the cut-matrices of* W' *and* W'' *have ranks equal to* 1.

Proof. Let π be the cut-matrix of W' in F. Since π has coefficients in $GF(2)$, rank $(\pi) = 1$ if and only if the nonnull lines of π are equal. Thus if W'^{+} is the index-set of the nonnull lines and if W''^{-} is the index-set of the nonnull entries of these lines, then $\delta W' = W'^{+} \times W''^{-}$. A similar argument can be used for the cut-matrix of W''. □

COROLLARY 6.6. *Local complementations do not change the splits of digraphs.*

We define a *split* of a diconnected Eulerian system $S = (L, V)$ as a bipartition $\{W', W''\}$ of V satisfying $|W'|, |W''| \geqq 2$ and $c(W') = c(W'') = 1$.

7. Decompositions. The reader is referred to § 1 for the definition of a decomposition $\{G', G''\}$ of a digraph G. He will easily verify that if G is diconnected, the same holds for G' and G''. The basic property, proved in [8], about digraph decompositions is the following one:

PROPOSITION 7.1. *If* $G \to \{G', G''\}$ *is a digraph decomposition then* $\{V(G'\backslash v), V(G''\backslash v)\}$ *is a split of* G. *Conversely for every split* $\{V', V''\}$ *of* G *and every marker* $v \notin V(G)$ *there exists precisely one decomposition* $G \to \{G', G''\}$. □

The decompositions of digraphs fall in the general theory of decomposition frames formulated by W. Cunningham and J. Edmonds [7], but this will no longer be true for the decompositions of Eulerian systems, which will now be defined, because of property 7.3. However these decompositions can be encompassed by generalized decomposition frames proposed by W. Cunningham [9]. This theory is not developed here but the idea is that there is a group G such that for each object N (the reader is referred to [7] for the terminology), each cell e of N and each $g \in G$ there is another object $N^{(e,g)}$ that satisfies some obvious axioms. In our case the group G will be equal to $GL(K)$ defined below.

Let $S' = (L', V')$ and $S'' = (L'', V'')$ be Eulerian systems, and let $W \subseteq V' \cup V''$. We shall always identify a vector $a \in K^{W}$ with a vector of $K^{V' \cup V''}$ by letting $a(v) = 0$ for every $v \in (V' \cup V'')\backslash W$. Thus L' and L'' will be considered as subspaces of $K^{V' \cup V''}$, and their sum $L' + L''$ can be defined as a subspace of $K^{V' \cup V''}$.

A pair of Eulerian systems $\{S', S''\}$, $S' = (L', V')$ and $S'' = (L'', V'')$, is called a (*simple*) *decomposition* of an Eulerian system $S = (L, V)$ if the three following conditions hold: (i) $V' \cap V''$ contains precisely one vertex v called the *marker*, (ii) $V = (V' \cup V'')\backslash\{v\}$, and (iii) $L = (L' + L'') \times V$.

PROPOSITION 7.2. *If* $\{S', S''\}$, $S' = (L', W' \cup \{v\})$ *and* $S'' = (L'', W'' \cup \{v\})$, *is a simple decomposition of an Eulerian system* $S = (L, V)$ *with the marker* v, *then*

(i) $\{W', W''\}$ *is a split of* S,

(ii) $L \times W' = L' \times W'$,

(iii) $L \times W'' = L'' \times W''$.

Proof. Since $L = (L' + L'') \times V$ and $W' \subseteq V$, we have

(a) $L' \times W' \subseteq L \times W'$,

which implies that

(b) $\dim (L' \times W') \leqq \dim (L \times W')$.

Since the dimension of K over $GF(2)$ is equal to 2, we have

(c) $\dim (L' \times W') \geqq \dim (L') - 2 = |W'| - 1$.

This implies that $c(W') \leqq 1$, and since S is diconnected the equality holds. Therefore inequalities are also equalities in (a)–(c), and this implies $c(W') = 1$ and (ii). Similarly, $c(W'') = 1$ and (iii) holds. □

It will be proved further that any split induces some simple decomposition. Here we show that a same split no longer induces a unique decomposition as was the case for digraphs.

Let $GL(K)$ be the group of the linear endomorphisms of the vector-space K. We notice that $GL(K)$ is the set of the permutations g of K satisfying $g(0) = 0$. For a vector $a \in K^V$, $v \in V$ and $g \in GL(K)$ let $a' = a^{(v,g)}$ denote the vector of K^V defined by $a'(w) = a(w)$ if $w \neq v$, and $a'(v) = g(a(v))$. For an Eulerian system $S = (L, V)$, $v \in V$ and $g \in GL(K)$ we let $S' = S^{(v,g)}$ denote the Eulerian system (L', V) defined by $L' = \{a^{(v,g)} : a \in L\}$.

PROPOSITION 7.3. *Two pairs $\{S'_1, S''_1\}$ and $\{S'_2, S''_2\}$ are simple decompositions of a same Eulerian system S with a same marker v and a same split if and only if there exists $g \in GL(K)$ such that $S'_2 = S'^{(v,g)}_1$ and $S''_2 = S''^{(v,g)}_1$.*

Proof. Let $S'_i = (L'_i, V')$ and $S''_i = (L''_i, V'')$ for $i = 1$ and 2. If $\{S'_1, S''_1\}$ and $\{S'_2, S''_2\}$ are decompositions of $S = (L, V)$ with the same marker v and the same split $\{W', W''\}$, following (7.2) we have

(1) $L'_1 \times W' = L'_2 \times W' = L \times W'$,

(2) $L''_1 \times W'' = L''_2 \times W'' = L \times W''$,

(3) $(L'_1 + L''_1) \times V = (L'_2 + L''_2) \times V = L$.

For each $x \in K$ and each $i \in \{1, 2\}$, let $L'_{ix} = \{a'_i \in L'_i : a'_i(v) = x\}$, and let L''_{ix} be similarly defined. We have

$$(L'_i + L''_i) \times V = \cup (L'_{ix} + L''_{ix} : x \in K).$$

We notice that $L'_{i0} = L'_i \times V'$ and $L''_{i0} = L''_i \times V''$. Therefore equalities (1)–(3) imply that $\{L'_{ix} + L''_{ix} : x \in K\}$ is the set of the cosets of $L \times W' + L \times W''$ in L. Therefore there exists $g \in GL(K)$ such that $L'_{2x} + L''_{2x} = L'_{1g(x)} + L''_{1g(x)}$ for every $x \in K$. If we make a projection onto $K^{W'}$ of the two members of the preceding equality, we obtain $L'_{2x} = L'_{1g(x)}$. Similarly we have $L''_{2x} = L''_{1g(x)}$. This implies the direct part of the statement. The converse is an easy verification. □

Digraphs and Eulerian systems are related by graphic presentations; thus in order to relate the decompositions of digraphs and those of Eulerian systems, we introduce the decompositions of graphic presentations. A pair of graphic presentations $\{(F', A', B'), (F'', A'', B'')\}$ is a (*simple*) *decomposition* of a graphic presentation (F, A, B) if the following conditions hold: (i) $\{F', F''\}$ is a decomposition of F with a marker v and a split $\{W', W''\}$; (ii) the restrictions of A and B to W' are respectively equal to those of A' and B' to W', and similarly the restrictions of A and B to W'' are respectively equal to those of A'' and B'' to W''; (iii) $A'(v) = B''(v)$ and $A''(v) = B'(v)$.

PROPOSITION 7.4. *Let $\{(F', A', B'), (F'', A'', B'')\}$ be a decomposition of a graphic presentation (F, A, B). If S, S', S'' are the Eulerian systems which are respectively induced by (F, A, B), (F', A', B') and (F'', A'', B'') then $\{S', S''\}$ is a decomposition of S.*

Proof. It presents no difficulty but it involves a lot of computation. Let $S = (L, V)$, $S' = (L', V')$, $S'' = (L'', V'')$, v be the marker of the decomposition of F. We let the reader verify the following lemma which is elementary algebra.

LEMMA. *Let x and y be two nonnull and distinct elements of K. Let E' be a base of L' and E'' be a base of L'' satisfying the following properties:*

There exist e'_x and e'_y in E', e''_x and e''_y in E'' satisfying $e'_x(v) = e''_x(v) = x$ and $e'_y(v) = e''_y(v) = y$;

For every $e' \in E' \backslash \{e'_x, e'_y\}$ and every $e'' \in E'' \backslash \{e''_x, e''_y\}$ we have $e'(v) = e''(v) = 0$.

Then the set $E = E' \backslash \{e'_x, e'_y\} \cup E'' \backslash \{e''_x, e''_y\} \cup \{e'_x + e''_x, e'_y + e''_y\}$ is a base of $(L' + L'') \times V$. □

Since F' is diconnected we can choose $w' \in V'$ such that $(w', v) \in A(F')$. Similarly we can choose $w'' \in V''$ such that $(w'', v) \in A(F'')$. We set

$$V'_1 = \{v' \in V' : (v', v) \notin A(F'), v' \neq v\},$$

$$V'_2 = \{v' \in V' : (v', v) \in A(F'), v' \neq w'\},$$

$$V''_1 = \{v'' \in V'' : (v'', v) \notin A(F''), v'' \neq v\},$$

$$V''_2 = \{v'' \in V'' : (v'', v) \in A(F''), v'' \neq w''\},$$

$$x = A'(v) = B''(v); \qquad y = A''(v) = B'(v),$$

$$n, n', n'' = \text{neighbourhood functions of } F, F', F''.$$

We notice that

$$V = V(F) = V'_1 \cup V'_2 \cup V''_1 \cup V''_2 \cup \{w', w''\}.$$

We consider the fundamental base of L' induced by the graphic presentation (F', A', B'), say

$$\{A'_{v'} = A'n'(v') + B'v' : v' \in V'\},$$

and we construct a new base $E' = \{a'_{v'} : v' \in V'\}$ defined by the formulas

$$a'_{v'} = A'_{v'} \quad \text{if } v' \in V'_1 \cup \{v, w'\}, \qquad a'_{v'} = A'_{v'} + A'_w \quad \text{if } v' \in V'_2.$$

Similarly we consider

$$\{A''_{v''} = A''n''(v'') + B''v'' : v'' \in V''\},$$

and we construct a new base $E'' = \{a''_{v''} : v'' \in V''\}$ of L'' by the formulas

$$a''_{v''} = A''_{v''} \quad \text{if } v'' \in V''_1 \cup \{v, w''\}, \qquad a''_{v''} = A''_{v''} + A''_{w''} \quad \text{if } v'' \in V''_2.$$

The bases E' and E'' satisfy the conditions of the lemma with $e'_x = a'_{w'}$, $e'_y = a'_v$, $e''_x = a''_v$, $e''_y = a''_{w''}$. Therefore, if $M = (L' + L'') \times V$, we have a base E of M defined by

(1) $E'_i = \{a'_{v'} : v' \in V'_i\}$, $\quad E''_i = \{a''_{v''} : v'' \in V''_i\}$,

$E = E'_1 \cup E'_2 \cup E''_1 \cup E''_2 \cup \{a'_{w'} + a''_v, a'_v + a''_{w''}\}$.

It remains to verify that E is also a base of L for proving that $S = (M, V)$. From the definition of the digraph composition it follows that the neighborhood function n is defined in terms of n' and n'' by the following formulas:

$$n(v') = n'(v') \quad \text{if } v' \in V'_1,$$

$$n(v') = n'(v') + n''(v) + v \quad \text{if } v' \in V'_2 \cup \{w'\},$$

$$n(v'') = n''(v'') \quad \text{if } v'' \in V'_2,$$

$$n(v'') = n''(v'') + n'(v) + v \quad \text{if } v'' \in V'_1 \cup \{w''\}.$$

Let us consider the fundamental base of L induced by the graphic presentation (F, A, B), say

$$\{A_w = An(w) + Bw : w \in V\}.$$

We derive from it a new base $H = \{\alpha_w : w \in V\}$ by the formulas

$$\alpha_{v'} = A_{v'} \quad \text{if } v' \in V'_1 \cup \{w'\},$$

$$\alpha_{v'} = A_{v'} + A_{w'} \quad \text{if } v' \in V'_2,$$

$$\alpha_{v''} = A_{v''} \quad \text{if } v'' \in V''_1 \cup \{w''\},$$

$$\alpha_{v''} = A_{v''} + A_{w''} \quad \text{if } v'' \in V''_2.$$

For $v' \in V'_1$ we have

$$\alpha_{v'} = A_{v'} = An(v') + Bv' = A'n'(v') + B'v' = A'_{v'} = a'_{v'}.$$

Therefore

(2) $E'_1 = (\alpha_{v'} : v' \in V'_1)$,

and similarly

(3) $E'_2 = (\alpha_{v''} : v'' \in V''_1)$.

For $v' \in V'_2 \cup \{w'\}$ we have

(4) $A_{v'} = A[n'(v') + n''(v) + v] + Bv'$.

This implies for $v' \in V'_2$

$$\alpha_{v'} = A_{v'} + A_{w'} = A[n'(v') + n'(w')] + B[v' + w'] = a'_{v'}.$$

Therefore

(5) $E'_2 = (\alpha_{v'} : v' \in V'_2)$,

and similarly

(6) $E''_2 = (\alpha_{v''} : v'' \in V''_2)$.

Let us consider again formula (4) for $v' = w'$. We can write

$$A_{w'} = A[n'(w') + v] + An''(v) + Bw'.$$

Since $n'(w') + v \subseteq V(F'\backslash v)$ and $n''(v) \subseteq V(F''\backslash v)$, we have

$$A_{w'} = A'[n'(w') + v] + A''n''(v) + B'w'$$

$$= A'n'(w') + A'v + A''n''(v) + B'w'.$$

The equality $A'(v) = B''(v)$ is equivalent to $A'v = B''v$. This implies that

$$A_{w'} = A'n(w') + B'w' + A''n''(v) + B''v.$$

Therefore

(7) $\alpha_{v'} = A_{w'} = a'_{w'} + a''_v$,

and similarly

(8) $\alpha_{v''} = A_{w''} = a''_{w''} + a'_v$.

Thus equalities (1)–(3), (5)–(8) imply $E = H$. □

COROLLARY 7.5. *For every split $\{W', W''\}$ of an Eulerian system $S = (L, V)$ and every marker $v \notin V$ we can construct a decomposition $\{S', S''\}$ such that $S' = \{L', W' \cup \{v\})$ and $S'' = (L'', W'' \cup \{v\})$.*

Proof. Let (F, A, B) be a graphic presentation of S. It follows from (6.1) and (6.5) that $\{W', W''\}$ is a split of F. Following (7.1) we construct a simple decomposition $\{F', F''\}$ of F with the marker v and the split $\{W', W''\}$. Let $V' = W' \cup \{v\}$ and $V'' = W'' \cup \{v\}$. We choose two distinct and nonnull elements x and y of K, and we define

$A' \in K^{V'}$ by $A'(v') = A(v')$ if $v' \in W'$ and $A'(v) = x$,

$B'' \in K^{V''}$ by $B''(v'') = B(v'')$ if $v'' \in W''$ and $B''(v) = x$,

$B' \in K^{V'}$ and $A'' \in K^{V''}$ similarly with $B'(v) = A''(v) = y$.

Thus we have a decomposition $\{(F', A', B'), (F'', A'', B'')\}$ of (F, A, B), and we apply the preceding result for constructing a decomposition of S. □

8. Semibrittle Eulerian system. A digraph F with vertex-set V (Eulerian system $S = (L, V)$) is *semibrittle* if $|V| \geqq 4$ and there exists an ordering $v_0, v_1, \cdots, v_{n-1}$ of V such that the splits of F are precisely the bipartitions $\{\{v_i, v_{i+1}, \cdots, v_{i+j-1}\}, \{v_{i+j}, \cdots, v_{i-1}\}\}$, where $0 \leqq i < n$, $1 < j < n - 1$ and subscripts are taken modulo n.

If F is a fundamental digraph of S, then the results of the preceding section imply that S is semibrittle if and only if F is semibrittle. Thus to characterize semibrittle Eulerian systems is equivalent to characterizing semibrittle digraphs. This has been done by Cunningham [7].

A *transitive tournament* is a digraph G such that for some ordering $v_1, v_2, \cdots, v_{n-1}$ of $V(G)$ we have $E(G) = \{(v_i, v_j) : 1 \leqq i < j \leqq n - 1\}$. A *circle of transitive tournaments* (CTT) is a digraph obtained from a sequence $T_0, T_1, \cdots, T_{k-1}$ of transitive tournaments which are of order at least 2 by identifying for each i the last vertex of T_i with the first vertex of T_{i+1} where the subscripts are taken modulo k. Each vertex produced by an identification is a *hinge*. More formally it will be convenient to define a CTT in the following way. If P is a circular permutation on V and $(x, y) \in V \times V$ then let $n(x, y, P) = \inf(j : j \geqq 0, y = P^j(x))$. The *open interval* $]x, y[_p$ is equal to

$$\{P^i(x) : 0 < i < n(x, y, P)\}.$$

For a nonempty subset $H \subseteq V$, the CTT denoted by $C(H, P)$ is the digraph F defined by $V(F) = V$, $E(F) = \{(x, y) : x \neq y,]x, y[_p \cap H = \varnothing\}$. The set of the hinges of $C(H, P)$ is H. For example $C(V, P)$ is the graph of the circular permutation P.

PROPOSITION 8.1 [7]. *A graph is semibrittle if and only if it is a CTT.* □

Since the fundamental digraphs of an Eulerian system are defined up to local complementation, we study now the effect of this operation on a CTT.

PROPOSITION 8.2. *The local complementation of $C(H, P)$ at a vertex v is equal to $C(H + v, P)$ if $H \neq \{v\}$, otherwise it is equal to $C(H, P^{-1})$.*

Proof. Let $F = C(H, P)$. Where

$$T = \{(x, y) : v \neq x \neq y \neq v, (x, v) \text{ and } (v, y) \in E(F)\},$$

we have $A(F*v) = A(F) + T$. We simplify the notation $]x, y[_p$ into $]x, y[$. Let us suppose first that $v \notin H$. We have

$$\begin{aligned} T &= \{(x, y) : v \neq x \neq y \neq v,]x, v[\cap H =]v, y[\cap H = \varnothing\} \\ &= \{(x, y) : v \neq x \neq y \neq v, (]x, v[\cup \{v\} \cup]v, y[) = \{v\}\}. \end{aligned}$$

In the above formula the subset $I =]x, v[\cup \{v\} \cup]v, y[$ must either be equal to the open interval $]x, y[$ or to V. The second case is excluded because $I \cap H = \{v\}$, and we have assumed that $v \notin H$. Therefore

$$T = \{(x, y) : x \neq y,]x, y[\cap H = \{v\}\}.$$

Let $H' = H + v$ and $F' = C(H', P)$. We have

$$\begin{aligned} A(F') &= \{(x, y) : x \neq y,]x, y[\cap H' = \varnothing\} \\ &= \{(x, y) : x \neq y,]x, y[\cap H = \varnothing \text{ or }]x, y[\cap H = \{v\}\} \\ &= \{(x, y) : (x, y) \in A(F) \text{ or } (x, y) \in T\}. \end{aligned}$$

In the above formula the "or" is obviously exclusive, and so we have $A(F') = A(F) + T$. Therefore F' is equal to $F*v$, which proves the statement when $v \notin H$. If $v \in H$ and $H \neq \{v\}$ the statement is proved because $F' = F*v$ is equivalent to $F = F'*v$.

Let us now consider the final case where $H = \{v\}$. For every vertex $x \neq v$, neither $]x, v[$ nor $]v, x[$ contains v. Therefore (v, x) and (x, v) are arcs of F. For every pair of vertices $\{x, y\}$ satisfying $v \neq x \neq y \neq v$, one and only one of the open intervals $]x, y[$ and $]y, x[$ contains v. Therefore one of the pairs (x, y) and (y, x) is a transitivity arc when the other one is a transitivity coarc. This implies that $F*v$ is obtained by reversing the arcs of F. But for every pair of vertices $\{x, y\}$ clearly we have $]x, y[_{p^{-1}} =]y, x[_p$. This implies that $C(H, P^{-1})$ is also obtained by reversing the arcs of $C(H, P)$. And so $C(H, P^{-1}) = F*v$. □

The following notation will be used. If P is a cyclic permutation and $s = +$ or $s = -$ then $P^s = P$ if $s = +$ and $P^s = P^{-1}$ if $s = -$. We prove naturally that $-s$ is the opposite of the sign s. We let $\mathbf{P}'(V) = \{(H, s) : \varnothing \neq H \subseteq V, s = + \text{ or } s = -\}$. For every $(H, s) \in \mathbf{P}'(V)$ and $v \in V$ we let $(H, s)*v = (H + v, s)$ if $H \neq \{v\}$ and $(H, s)*v = (H, -s)$ if $H = \{v\}$.

PROPOSITION 8.3. $\mathbf{P}'(V) = \{(V, +)*m : m \in V^*\}$.

Proof. If $v_1, v_2, \cdots, v_n$ is an ordering of V it is easy to verify that:

$$(V,+)*v_1v_2 \cdots v_k = (V\backslash\{v_1, v_2, \cdots, v_k\}, +) \quad \text{if } k<n,$$

$$(V,+)*v_1v_2 \cdots v_nv_{n-1} \cdots v_k = (\{v_n, v_{n-1}, \cdots, v_k\}, -) \quad \text{if } k \leqq n.$$

The result follows. □

PROPOSITION 8.4. *For every semibrittle Eulerian system $S = (L, V)$ we can find a cyclic permutation P on V and a pair $\{X^+, X^-\}$ of supplementary vectors of K^V such that, if we let $X = X^+ + X^-$, the set of the graphic presentations of S is*

$$\{GP(H, s) : (H, s) \in \mathbf{P}'(V)\}$$

where

$$GP(H, s) = (F(H, s), A(H, s), B(H, s)),$$

$$F(H, s) = C(H, P^s),$$

$$A(H, s) = X + X^sH,$$

$$B(H, s) = X^s \quad \text{if } |H| > 1, \quad B(H, s) = X^s + X^{-s}H \quad \text{if } |H| = 1.$$

Proof. If we consider a fundamental digraph of S, it is a CTT. Let it be equal to $C(H, P)$. Let $v_1v_2 \cdots v_k$ be an ordering of $V\backslash H$. Proposition 8.2 implies that $C(H, P)*v_1v_2 \cdots v_k = C(V, P)$. Therefore $C(V, P)$ is a fundamental digraph of S, and we can consider a graphic presentation of S of the form $(C(V, P), X, X^+)$. We prove that the statement holds with the values of P, X, X^+ which are so determined and $X^- = X + X^+$. First we notice that $GP(V, +) = (C(V, P), X, X^+)$ is a graphic presentation of S. Let us show first that the statement will be proved if

(i) $(GP(H, s))*v = GP((H, s)*v)$, $(H, s) \in \mathbf{P}'(V)$, $v \in V$.

Indeed if Z is the set of all the graphic presentations of S, we have

$$\begin{aligned} Z &= \{(GP(V, +))*m : m \in V^*\} && \text{by (5.2)}, \\ &= \{GP((V, +)*m) : m \in V^*\} && \text{if (i) is true}, \\ &= \{GP(H, s) : (H, s) \in \mathbf{P}'(V)\} && \text{by (8.3)}. \end{aligned}$$

From now on the verification of (i) is based on (4.1). First we define notation consistent with (4.1):

$$GP = GP(H, s), \quad F = F(H, s), \quad A = X + X^sH,$$

$$B = X^s \quad \text{if } |H| > 1 \quad \text{and} \quad B = X^s + X^{-s}H \quad \text{if } |H| = 1,$$

$$n = \text{neighborhood function of } F,$$

$$F' = F*v, \quad A' = A + Bv, \quad B' = B + A[n(v) \cap n^{-1}(v)].$$

Verifying (i) amounts to verifying successively

(ii) $F' = F((H, s)*v)$,

(iii) $A' = A((H, s)*v)$,

(iv) $B' = B((H, s)*v)$.

First we notice that (ii) is directly implied by (8.2). Let us verify (iii). We have

$$A' = A + Bv = (X + X^sH) + Bv.$$

If $|H| > 1$ this implies that

$$A' = (X + X^sH) + X^sv = X + X^s[H + v].$$

If $|H| = 1$ and $H = \{w\} \neq \{v\}$ this implies that

$$\begin{aligned} A' &= (X + X^s w) + (X^s + X^{-s} w)v \\ &= X + X^s w + X^s v = X + X^s[H + v]. \end{aligned}$$

If $H = \{v\}$ this implies that

$$\begin{aligned} A' &= (X + X^s v) + (X^s + X^{-s} v)v \\ &= X + X^s v + X^s v + X^{-s} v = X + X^{-s} H. \end{aligned}$$

In each case we have well $A' = A((H, s) * v)$. In order to verify (iv) we compute first the value of $n(v) \cap n^{-1}(v)$. If we simplify the notation $]x, y[_{P^s}$ into $]x, y[$, we have

$$n(v) \cap n^{-1}(v) = \{w \neq v :]v, w[\cap H =]w, v[\cap H = \varnothing\}.$$

For any two distinct vertices v and w, we have

$$]v, w[\cup]w, v[= V \backslash \{v, w\}.$$

Therefore

$$n(v) \cap n^{-1}(v) = \{w \neq v : H \subseteq \{v, w\}\}.$$

If $|H| > 2$ or $|H| = 2$ and $v \notin H$, we have $n(v) \cap n^{-1}(v) = \varnothing$. This implies that

$$B' = B = X^s.$$

From now on it will be convenient to express A as $X^s + X^{-s} + X^s H$. If $|H| = 2$ and $v \in H$, let $H = \{v, w\}$. We have $n(v) \cap n^{-1}(v) = \{w\}$. Therefore

$$\begin{aligned} B' &= B + Aw = X^s + (X^s + X^{-s} + X^s[v + w])w \\ &= X^s + X^s w + X^{-s} w + X^s w = X^s + X^{-s} w. \end{aligned}$$

If $|H| = 1$ and $H = \{w\} \neq \{v\}$, we have $n(v) \cap n^{-1}(v) = \{w\}$. This implies that

$$\begin{aligned} B' &= B + Aw = (X^s + X^{-s} w) + (X^s + X^{-s} + X^s w)w \\ &= X^s + X^{-s} w + X^s w + X^{-s} w + X^s w = X^s. \end{aligned}$$

If $H = \{w\}$, we have $n(v) \cap n^{-1}(v) = V \backslash v$. Let $V' = V \backslash v$. We have

$$\begin{aligned} B' &= B + AV' = (X^s + X^{-s} v) + (X^s + X^{-s} + X^{-s} v)V' \\ &= X^s + X^{-s} v + X^s V' + X^{-s} V' = X^{-s} + X^s v. \end{aligned}$$

In each case we have well $B' = B((H, s) * v)$. □

Let us apply the preceding proposition when $|V| = 3$. We let $V = \{1, 2, 3\}$ and $K = \{0, x, y, z\}$. Every complete vector $a \in K^V$ will be represented by the sequence $a(1)a(2)a(3)$. If we have $X^+ = xxx$ and $X^- = yyy$ then the set of the Eulerian vectors is equal to

$$\{yyy, yyz, yzy, zyy, yzz, zyz, zzy, zzx, zxz, xzz, zxx, xzx, xxz, xxx\}.$$

We notice that the Hamming distance between $a_1 = yyy$ and $a_2 = xxx$ is equal to 3 when a shortest word m satisfying $a_2 = a_1 * m$ is, for example, $m = 12321$. Thus the upper bound of the accessibility property 5.1 is reached. We shall prove in a future article that it is in some sense a characteristic property of semibrittle Eulerian systems.

COROLLARY 8.5. *A semibrittle Eulerian system* $S = (L, V)$ *is of index* 1 *if* $|V| \geqq 3$.

Proof. Following (8.4) each fundamental digraph $F(H, s)$ is equal to $C(H, P^s)$ for $(H, s) \in \mathbf{P}'(V)$. These CTT's are pairwise distinct. □

If $|V| = 2$ there are six graphic presentations, and all the fundamental graphs are reduced to a single edge.

9. An algorithm for searching for a split of a digraph. Such an algorithm has been devised by W. Cunningham [8]. If the given digraph G has n vertices, then the time-complexity of Cunningham's algorithm is $O(n^4)$. In this section we describe an algorithm with an $O(n^3)$ time-complexity.

For $W \subseteq V(G)$ and $s \in V(G)\backslash W$ we say that a split $\{V', V''\}$ of G *separates* W from s if $W \subseteq V'$ and $s \in V''$. It has been proved by Cunningham, [8, Problem 1 and Thm. 13] that such a split can be found in $O(n^2)$ time if we know two arcs (v, s) and (s, w) of G such that $v, w \in W$. Throughout this section we identify by SEPAR (G, W, v, s, w) such a procedure. Where $m = |A(G)|$ we notice that in Cunningham's algorithm G is represented in $O(m)$ space by keeping the in-list and the out-list of each vertex of G. This will no longer be possible with the present algorithm which will perform local complementations. Thus G will be represented by its adjacency matrix which takes $O(n^2)$ space. However this does not change the time-complexity of SEPAR.

The variables used by the algorithm are a subset $W \subseteq V$, a stack S of vertices of G, and a vertex $w \in W$. We initialize with $W = \{w\}$ where w is an arbitrary vertex and S is empty. The properties satisfied by the variables W, S, w each time they are read are the following:

(i) Every split $\{V', V''\}$ satisfies either $W \subseteq V'$ or $W \subseteq V''$;

(ii) If S is nonempty and if s is the vertex at the top of S then $s \notin W$, $(s, w) \in A(G)$, and there exists a vertex $v \in W$ such that $(v, s) \in A(G)$.

If S is empty and $|W| \leqq n - 2$ then the algorithm calls a procedure FILLSTACK which places some vertices in S and determines w such that (ii) holds. If S is nonempty the algorithm removes the vertex s at the top of S, searches for v satisfying (ii) and calls SEPAR (G, W, v, s, w). If SEPAR returns with a split which separates W from s then the algorithm stops, otherwise it lets $W = W + s$. The property (i) is again satisfied and the algorithm iterates. If the graph G is found to be prime then the time-complexity involved by the successive calls of SEPAR is $O(n^3)$ and it is $O(n^2)$ for the successive searches for $v \in W$ satisfying (ii).

Throughout the end of this section we define a *path* of a digraph H as a sequence of pairwise distinct vertices $\Gamma = z_0, z_1, \cdots, z_m$ such that (z_i, z_{i+1}) is an arc of H for every $i = 0, 1, \cdots, m - 1$. A *chord* of Γ is any arc (z_i, z_j) such that $0 \leqq i < j \leqq 1$ and $j - i > 1$. We call (z_0, z_m) the *extremal chord* of Γ. We notice that arcs (z_i, z_j) with $i > j$ are not considered to be chords of Γ.

The first step of FILLSTACK is to construct in G a path $s_0, s_1, \cdots, s_q, s_{q+1}$ without nonextremal chord and such that $q > 0$, s_0 and $s_{q+1} \in W$, $s_1, s_2, \cdots, s_q \notin W$. This can be done in $O(m)$ time by searching for an arc (s_0, α) such that $s_0 \in W$ and $\alpha \notin W$ (such an arc exists because G is diconnected), constructing a shortest path Γ from α to W by means of a classical algorithm, taking for s_1 the last vertex of Γ such that $(s_0, s_1) \in A(G)$ and defining $s_1, s_2, \cdots, s_{q+1}$ as the subpath of Γ starting at s_1. The second step pushes $s_1, s_2, \cdots, s_q$ into S (so that the main algorithm will use successively $s_q, s_{q-1}, \cdots, s_1$), lets $w = s_{q+1}$, and performs successive local complementations at $s_1, s_2, \cdots, s_q$, which does not change the splits by (6.6). The time-complexity of FILLSTACK is $qO(n^2)$ for pushing q vertices into S, so that the overall complexity involved by the calls to FILLSTACK when G is prime is again $O(n^3)$.

In order to prove that FILLSTACK returns S and w satisfying (ii) we let $G_0 = G$, $G_i = G * s_1 s_2 \cdots s_i$ for each $i = 1, 2, \cdots, q$, and we verify by induction that the following properties hold for $i = 0, 1, \cdots, q$:

(a) $s_0, s_{i+1}, \cdots, s_{q+1}$ is a path of G_i without nonextremal chord except possibly when $i = q$;

(b) $(s_j, s_{i+1}) \in A(G_i)$ $(1 \leqq j \leqq i)$;

(c) either $(s_0, s_j) \in A(G_i)$ or there exists $j' > j$ such that $(s_{j'}, s_j) \in A(G_i)$ $(0 < j \leqq i)$.

We notice that $G_q = G$ at the return of FILLSTACK, and we show that properties (b)–(c) for $i = q$ imply (ii). Any s removed from the top of S is equal to some s_j $(1 \leqq j \leqq q)$ and we have $w = s_{q+1}$, so that (b) implies $(s, w) \in A(G)$. Moreover the vertices $s_q, s_{q-1}, \cdots, s_{j+1}$ belong to W when s_j is removed from the top of S and (c) implies that either (s_0, s) or $(s_{j'}, s)$, $j' > j$, is an arc of G, so that we find actually a vertex v (either s_0 or $s_{j'}$) in W satisfying $(v, s) \in A(G)$.

Properties (a)–(c) are clearly satisfied for $i = 0$. We assume that they hold for some $i \geqq 0$. Property (a) holds for $i + 1$: First we verify that $(s_0, s_{i+2}) \in A(G_{i+1})$ except possibly when $i + 1 = q$. Indeed this arc does not appear in $A(G_i)$ since otherwise it would be a nonextremal chord of $s_0, s_{i+1}, \cdots, s_{q+1}$, and it appears in $A(G_{i+1})$ because (s_0, s_{i+1}), $(s_{i+1}, s_{i+2}) \in A(G_i)$. Now we verify that $(s_k, s_{k+1}) \in A(G_i)$ for $i + 1 \leqq k \leqq q$. Indeed it is an arc of G_i, and for disappearing in G_{i+1} it would be necessary that $(s_k, s_{i+1}) \in A(G_i)$ and $(s_{i+1}, s_{k+1}) \in A(G_i)$, which is possible for the second arc only if $k = i + 1$ but then the first arc cannot be in G_i. Thus $s_0, s_{i+2}, \cdots, s_{q+1}$ is a path of G_{i+1}. To verify that it has no nonextremal chord is similar, and (b) is also similarly verified for $i + 1$. Let us verify (c) for $i + 1$. First we notice that (s_0, s_{i+1}) is an arc of G_{i+1} because it is already an arc of G_i (Property (a) for i) and no arc of G_i incident to s_{i+1} is modified by the local complementation at s_{i+1}. Thus let us consider any j satisfying $0 < j < i + 1$. In G_i we have either the arc (s_0, s_j) or $(s_{j'}, s_j)$ with some $j' > j$. If this arc is not removed from G_i after the local complementation at s_{i+1}, all is done. Otherwise the arc (s_{i+1}, s_j) must exist, and so we can let $j' = i + 1$.

Finally we have proved the following:

PROPOSITION 9.1. *There exists an algorithm in $O(n^3)$ time and $O(n^2)$ space for finding an eventual split of a diconnected graph G of order n.* □

Acknowledgment. Conversations and correspondence with William Cunningham have helped me in developing this work on decompositions. I am grateful to him.

REFERENCES

[1] J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, North-Holland, New York, 1981.

[2] A. BOUCHET, *Isotropic systems*, European J. Combin., to appear.

[3] ———, *Graphic presentations of isotropic systems*, submitted.

[4] ———, *Un algorithme polynomial pour reconnaître les graphes d'alternance*, C. R. Acad. Sci. Paris Sér. I Math., 300 (1985), pp. 569–572.

[5] ———, *Reducing prime graphs and recognizing circle graphs*, Combinatorica, to appear.

[6] ———, *Connectivity of isotropic systems*, submitted.

[7] W. H. CUNNINGHAM AND J. EDMONDS, *A combinatorial decomposition theory*, Canad. J. Math., 32 (1980), pp. 734–765.

[8] W. H. CUNNINGHAM, *Decomposition of directed graphs*, this Journal, 3 (1982), pp. 214–228.

[9] ———, personal communication.

[10] C. P. GABOR, WEN-LIAN HSU AND K. J. SUPOWIT, *Recognizing circle graphs in polynomial time*, Proc. IEEE Symposium on the Foundations of Computer Science, to appear.

[11] W. NAJI, *Graphes de cordes. Caractérisation et reconnaissance*, Discrete Math., 54 (1985), pp. 329–337.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

004

PERMUTATIONS WITH RESTRICTED DISPLACEMENT*

HENRY BEKER† AND CHRIS MITCHELL‡

Abstract. The permanent of an n by n (0, 1) circulant matrix is known to be equal to the number of permutations on n objects satisfying certain positional restrictions. The size of this number is of major importance for the design of certain analogue speech scramblers, as well as being a generalisation of certain "classical" enumeration problems. In this paper a new method is given for evaluating this permanent, which gives as corollaries many of the previously known results. The analogue speech scrambling scheme is also used to motivate a second enumeration problem, about which little seems to be known.

Key words. speech security, cryptography, permanents, analogue speech scramblers, derangements, menages

AMS(MOS) subject classifications. 05A15, 15A15, 94A60

1. Introduction. In this paper we consider a permutation enumeration problem that is of interest for two main reasons. First, it is a classical combinatorial problem, the study of certain cases of which goes back to the last century. Second, it is of considerable practical significance in the field of cryptography, in particular to the designers of time element speech scramblers.

The set of permutations that concern us here we call $A(n, k)$, where $1 \leqq k \leqq n$, and $A(n, k)$ contains permutations of $\{1, 2, \cdots, n\}$. More formally, we define

$$A(n,k)=\{\pi\in S_n : \overline{i\pi}\in\{\bar{i},\overline{i+1}, \cdots, \overline{i+k-1}\} \text{ for every } i\}$$

where the $\bar{\ }$ indicates the equivalence class modulo n.

In a combinatorial context, $|A(n, k)|$ has been studied under many guises; in particular note that evaluating $|A(n, k)|$ for $k = n - 1$ is the "problème des rencontres," and for $k = n - 2$ is the "problème des ménages." In addition, $|A(n, k)|$ is equal to the permanent of a certain (0, 1) n by n matrix. For a study of results in this context the reader is referred to Minc's unique book [11].

$|A(n, k)|$ is also of considerable practical significance because it is equal to the number of different "scrambling patterns" that can be used in a certain type of time element scrambling speech encryption device. For a more general introduction to this type of application see [1]-[4] and [12].

In this paper we consider a new approach to the evaluation of $|A(n, k)|$ which gives a direct method of computing it as the sum of the traces of the nth powers of $[(k-1)/2]$ matrices containing only zeros and ones. This new approach gives as immediate corollaries both the recurrence relations of Metropolis, Stein and Stein [9], and a number of previously well-known results.

Although the computational method requires a prohibitively large amount of computer storage for practical use in computing $|A(n, k)|$ for values of k much in excess of 12, it enables the computation of $|A(n, k)|$ for values of n and k not previously accessible. In particular, since the running time of the computation method is polynomial in n for fixed k, values of $|A(n, k)|$ can be directly computed for relatively large values of n given that k is sufficiently small. It is not surprising that computing $|A(n, k)|$ seems a difficult problem, since Valiant ([17] and [18]) has shown that evaluating the permanent of a (0, 1) matrix is a #P-complete problem (see also Garey and Johnson [5]).

* Received by the editors September 3, 1985; accepted for publication (in revised form) September 11, 1986.

† Racal-Guardata Ltd., Fleet, Hampshire GU13 8BU, England.

‡ Hewlett-Packard Ltd., Stoke Gifford, Bristol BS12 6QZ, England.

2. The combinatorial problem. Throughout this paper we will write $a(n, k)$ for the cardinality of the set $A(n, k)$, i.e.,

$$a(n,k) = |\{\pi \in S_n : \overline{i\pi} \in \{\bar{i}, \overline{i+1}, \cdots, \overline{i+k-1}\} \text{ for every } i\}|.$$

As in [11], $a(n, k)$ can also be defined as the permanent of the n by n matrix $Q(n, k)$, where

$$Q(n,k) = \sum_{i=0}^{k-1} P^i$$

and where P denotes the n by n permutation matrix with a one in positions (1, 2), (2, 3), $\cdots$, $(n-1, n)$, $(n, 1)$.

Explicit formulae for $a(n, k)$ have only been derived for values of k either near 0 or near n. We first consider the known results for which k is close to n.

Clearly $A(n, n) = S_n$, and hence $a(n, n) = n!$. As noted above, evaluating $a(n, n-1)$ is the well-known "problème des rencontres," and $a(n, n-1)$ is equal to the number of elements of S_n having no fixed point. The number $a(n, n-1)$ is often written as D_n (the "derangements number"), and is discussed in many combinatorial texts, see for example [7, pp. 541–542]. Similarly, evaluating $a(n, n-2)$ is also a well-known problem, commonly called the "problème des ménages." The solution for both these problems goes back to the last century; according to [6], a formula for $a(n, n-2)$ was obtained by Cayley and Muir in 1878.

The evaluation of $a(n, n-3)$ was first considered in [14], which contains no explicit formula but does give an asymptotic result. Yamamoto [20] considered the same problem, but it was Moser [13], who first produced an explicit formula for $a(n, n-3)$, which is, however, rather complex. Whitehead [19] has more recently considered the problem of evaluating $a(n, n-4)$.

In summary we have the following.

Result 2.1. (i) $a(n, n) = n!$, $n \geqq 1$.

(ii) $a(n, n-1) = n! \sum_{i=0}^{n} (-1)^i/i!$, $n \geqq 2$.

(iii) $a(n, n-2) = \sum_{i=0}^{n} (-1)^i.2n. \binom{2n-i}{i}.(n-i)!/(2n-i)$, $n \geqq 3$ (Touchard (see [6] and [15])).

For Moser's formula for $a(n, n-3)$ the interested reader is referred to [13].

Second, we consider results for small k. The cases $k = 1$ and 2 are trivial, and simple recurrence relations for $k = 3$ and 4 have been derived by Minc (see [10] or [11]). In addition, Metropolis, Stein and Stein [9] give recursion formulae for $a(n, k)$ for $k \leqq 9$.

In summary we have the following.

Result 2.2. (i) $a(n, 1) = 1$, $n \geqq 1$.

(ii) $a(n, 2) = 2$, $n \geqq 2$.

(iii) $a(n, 3) = a(n-1, 3) + a(n-2, 3) - 2$, $n \geqq 5$, $a(3, 3) = 6$ and $a(4, 3) = 9$.

(iv) $a(n, 4) = a(n-1, 4) + a(n-2, 4) + a(n-3, 4) - 4$, $n \geqq 7$, $a(4, 4) = 24$, $a(5, 4) = 44$ and $a(6, 4) = 80$.

The recursion formulae of [9] for $5 \leqq k \leqq 9$ are much more complex.

3. Time element speech scramblers. The practical application of permutations in $A(n, k)$ is in a certain kind of speech scrambler called a *time element scrambler.* There are a variety of types of time element scrambler systems, but they all employ the same general principle. The technique relies on the scrambler "recording" segments of speech, and then transmitting these segments in a different order.

More specifically, in a conventional so-called *hopping window* time element scrambler, the analogue speech signal is first divided into equal time periods called *frames.* Each frame is then further subdivided into a fixed number n of small equal time periods called *segments*, where the length of a segment would typically be of the order of 25–50 milliseconds. The scrambling is then achieved by transmitting the segments within a frame in a permuted order. At the receiver the inverse permutation is used to recover the original speech.

A typical system for $n = 8$ is illustrated in Fig. 1 below. For a more detailed discussion of the design considerations for such a device, such as the choices for n, the segment length and the selection of permutations to use in the scrambler, the reader is referred to [3]. The important thing to note here is that the system delay for such a device will be $2nT$ seconds, given that T is the segment length.

Thus, if T is, say, 50 milliseconds, and if $n = 8$, then the system delay will be 0.8 seconds, which is large enough to be noticeable. For larger n and T this delay will become unacceptably long, and yet, if $n = 8$, the total number of available permutations is only $8! = 40320$. So a problem can arise over choosing n sufficiently large to give a wide enough choice of enciphering permutations, and choosing n small enough to make the system delay acceptably short.

The idea of *sliding window* time element scramblers is to reduce the inherent time delay of the system, whilst at the same time increasing the number of possible scrambling patterns that can be used. There are a number of different types of sliding window systems, and for a description of some of these see [13] and [12]; we consider here one particular type, which we call *overlapping frame* sliding window time element scrambling, chosen for its ease of implementation.

As in a straightforward time element scrambler, the speech is again divided into frames of n segments, where each segment is T seconds long. However, we restrict ourselves to using a special subset of permutations from S_n, and we use these permutations in a slightly different way.

We first choose an integer k less than n. As we shall see, the choice of k directly affects the total system delay, which is equal to $(k + 1)T$ seconds. Thus if $k = 16$ and $T = 30$ milliseconds then the system delay would be 0.51 seconds. Note also that the system delay is independent of the choice of n.

Having fixed k, we then restrict our choice for scrambling permutations from S_n to those permutations π satisfying:

$$\overline{i\pi} \in \{\overline{\mathrm{i}-1}, \overline{i-2}, \cdots, \overline{i-k}\} \quad \text{for each } i \ (1 \leqq i \leqq n)$$

where $\bar{i}$ denotes the residue class of i modulo n. The idea is that at time t the segment spoken at time s is transmitted, where $\bar{s} = \overline{t\pi}$ and $1 \leqq t - s \leqq k$; this is possible because π satisfies the above property.

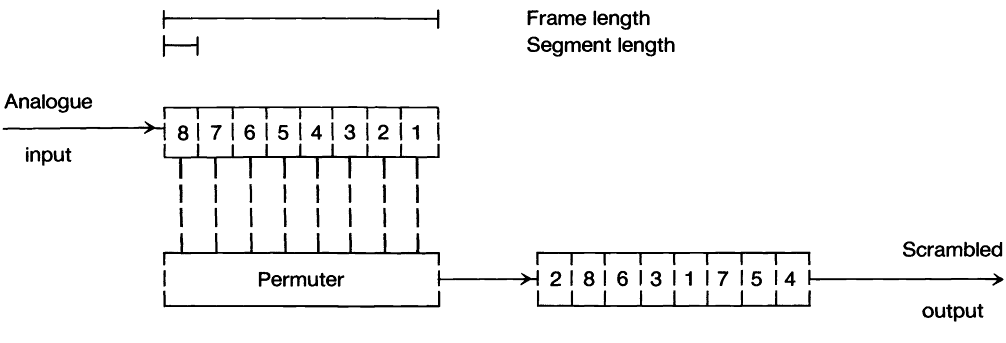


FIG. 1. *Hopping window time element scrambling.*

In Fig. 2, the use of such a permutation is illustrated for a system having $n = 8$, $k = 3$ and

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 1 & 8 & 3 & 2 & 5 & 4 & 7 \end{pmatrix}.$$

In the figure we have used different letters to distinguish between frames, so that A1, A2, $\cdots$, A8 are used to denote the eight segments of the first frame, B1, B2, $\cdots$, B8 denote the segments of the second frame, and so on.

Because of the condition imposed on the permutation, we know that each segment will be transmitted at most $3T$ seconds after it has been spoken, and hence the receiver can output the recovered descrambled speech signal $4T$ seconds after it has been input to the transmitting device. In general, each segment will be transmitted within kT seconds and thus the total system delay will be $(k+1)T$ seconds. This assumes that each segment must spend at least T seconds in both the transmitting and receiving devices.

As we have stated above, the only permutations that are usable in this type of sliding window time element scrambler are those permutations $\pi \in S_n$ satisfying:

$$\overline{i\pi} \in \{\overline{i-1}, \overline{i-2}, \cdots, \overline{i-k}\} \quad \text{for every } i.$$

In this paper we are concerned with the problem of enumerating these permutations, which is obviously a problem of considerable practical cryptographic significance.

As in § 1 above, we thus define:

$$A^*(n,k) = \{\pi \in S_n : \overline{i\pi} \in \{\overline{i-1}, \overline{i-2}, \cdots, \overline{i-k}\} \text{ for every } i\}$$

and we are interested in $a(n, k) = |A^*(n, k)|$.

For the purposes of the theory which follows it is easier to consider the set:

$$A(n,k) = \{\pi \in S_n : \overline{i\pi} \in \{\overline{i}, \overline{i+1}, \cdots, \overline{i+k-1}\} \text{ for every } i\}$$

and it is clear that $a(n, k) = |A(n, k)|$.

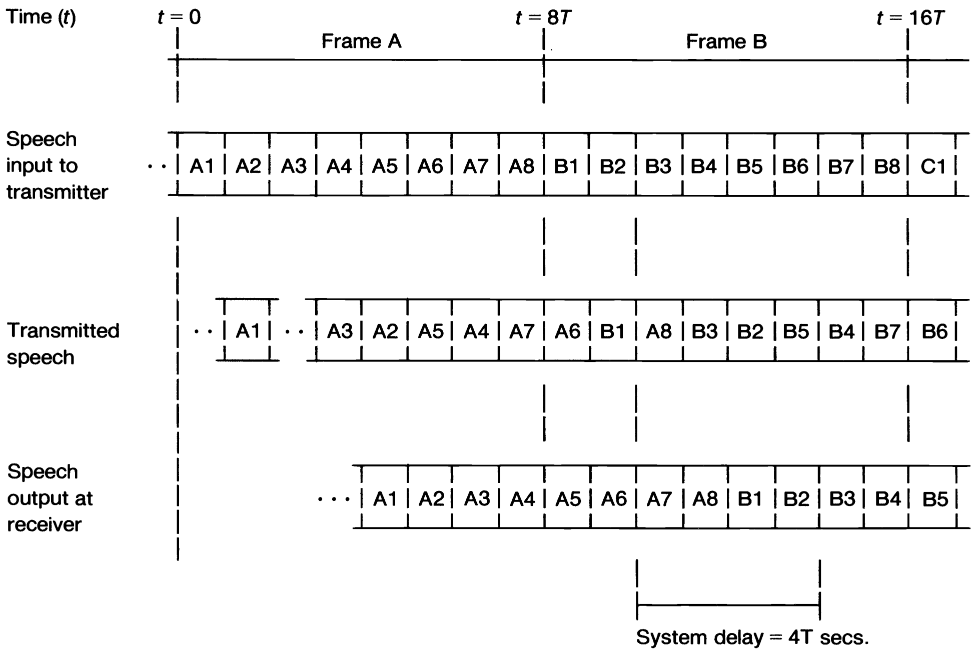

Time (t)
t = 0
t = 8T
t = 16T
Frame A
Frame B
Speech input to transmitter
·· A1 A2 A3 A4 A5 A6 A7 A8 B1 B2 B3 B4 B5 B6 B7 B8 C1 ·
Transmitted speech
·· A1 ·· A3 A2 A5 A4 A7 A6 B1 A8 B3 B2 B5 B4 B7 B6 ·
Speech output at receiver
··· A1 A2 A3 A4 A5 A6 A7 A8 B1 B2 B3 B4 B5 ·
System delay = 4T secs.

FIG. 2. *Overlapping frame sliding window time element scrambling.*

4. The main result and some corollaries. In this section we state the main results of this paper without proof; the proofs of all the results given here can be found in the next section.

Before stating the main theorem we first need a little notation. If $\pi \in S_n$ and $i \in \{1, 2, \cdots, n\}$, then define

$$X_k(\pi, i) = \{j\pi, j \in \{\overline{i}, \overline{i+1}, \cdots, \overline{i+k-2}\} : j\pi \in \{\overline{j}, \overline{j+1}, \cdots, \overline{i+k-2}\}\}.$$

Clearly by definition, $0 \leqq |X_k(\pi, i)| \leqq k - 1$.

We now state the following result, which is of fundamental importance.

LEMMA 4.1. *If $\pi \in A(n, k)$, then there exists an integer r, $r \in \{0, 1, \cdots, k-1\}$, such that*

$$|X_k(\pi, i)| = r \quad \textit{for every } i \in \{1, 2, \cdots, n\}.$$

Because of this result, we make the following definition:

$$A(n, k, r) = \{\pi \in A(n, k) : |X_k(\pi, i)| = r \text{ for every } i\}, \qquad 0 \leqq r \leqq k-1.$$

By Lemma 4.1 it is clear that $A(n, k)$ is equal to the disjoint union of the $A(n, k, r)$'s for r satisfying $0 \leqq r \leqq k - 1$, and hence, if we let $a(n, k, r) = |A(n, k, r)|$, then we have the following lemma.

LEMMA 4.2.

$$a(n, k) = \sum_{r=0}^{k-1} a(n, k, r).$$

In fact, in order to compute $a(n, k)$ using this lemma it is only necessary to compute $a(n, k, r)$ for r satisfying $1 \leqq r \leqq [(k-1)/2]$, since we also have the following.

LEMMA 4.3. (i) $a(n, k, r) = a(n, k, k-1-r)$, $0 \leqq r \leqq k-1 \leqq n-1$.

(ii) $a(n, k, 0) = a(n, k, k-1) = 1$, $0 \leqq k-1 \leqq n-1$.

Now suppose k and r are integers satisfying $0 \leqq r \leqq k - 1$, and let $t = \binom{k-1}{r}$. Label the t distinct r-subsets of $\{0, -1, \cdots, -k+2\}$: $R_1, R_2, \cdots, R_t$, and let

$$R_i^* = \{j + 1 : j \in R_i - \{0\}\},$$

for every $i \in \{1, 2, \cdots, t\}$.

Then define the t by t matrix $H(k, r) = (h_{ij})$ by

$$h_{ij} = \begin{bmatrix} 1 & \text{if } R_i^* \text{ is a subset of } R_j \\ 0 & \text{otherwise} \end{bmatrix}.$$

We can now state the main result.

THEOREM 4.4. $a(n, k, r) = \operatorname{Trace}(H(k, r)^n)$.

This result, in combination with Lemmas 4.2 and 4.3, provides a direct method for computing $a(n, k)$. Unfortunately, for k much larger than 12, $H(k, r)$ becomes extremely large, and the method is unusable because of the computer storage requirements. However, for fixed $k \leqq 12$, $a(n, k)$ can be computed for large n without much difficulty.

Furthermore, this theorem has as an immediate corollary, the recurrence relations of [9]. By the Cayley–Hamilton Theorem, $H(k, r)$ satisfies its own characteristic equation, and hence, since Trace is a linear function, $a(n, k, r)$ satisfies the characteristic equation of $H(k, r)$. The following corollary results.

COROLLARY 4.5 (Metropolis, Stein and Stein). *Suppose that*

$$\det (H(k, r) - xI) = \sum_{i=0}^{t} c_i x^i.$$

Then

$$\sum_{i=0}^{t} c_i a(n+i,k,r)=0.$$

Moreover, since the Trace of a matrix is equal to the sum of its eigenvalues, we immediately have the following corollary.

COROLLARY 4.6. *Suppose* $q_1, q_2, \cdots, q_t$ *are the eigenvalues of* $H(k, r)$. *Then*

(i) $a(n, k, r) = \sum_{i=1}^{t} q_i^n$.

(ii) *If* $s \in \{1, 2, \cdots, t\}$ *satisfies* $|q_s| > |q_i|$ *for every* $i \in \{1, 2, \cdots, t\} - \{s\}$, *then* $a(n, k, r)/(q_s)^n$ *tends to* 1 *as* n *tends to infinity.*

Note that (ii) above has as an immediate corollary a conjecture of [9], namely that $a(n+1, k, r)/a(n, k, r)$ tends to q_s as n tends to infinity; [9, Table IV] lists the maximal eigenvalues of $H(k, r)$ for $k = 4, 5, 6, 7, 8, 9$ and all r satisfying $1 \leqq r \leqq [(k-1)/2]$. Corollary 4.6 (ii) also implies that, for a fixed k, $a(n, k)$ is asymptotic to $(q_{\max})^n$, where $q_{\max}$ is the maximum value from the set of eigenvalues for all the matrices $H(k, r)$, $0 \leqq r \leqq [(k-1)/2]$.

Although $a(n, k, r)$ has meaning only if $n \geqq k$, $H(k, r)^n$ exists for every $n \geqq 1$. We can thus define $a(n, k, r)$ to be Trace $(H(k, r)^n)$ for every n satisfying $1 \leqq n \leqq k-1$. These values of $a(n, k, r)$ will clearly satisfy Corollary 4.5, and so they can be used as initial values for the recurrence relation.

The case $r = 1$ is an especially tractable one, and we give below a complete solution for this case. The first of the results is given in [9], but the second seems to be previously unknown.

Result 4.7. (Metropolis, Stein and Stein).

$$a(n+k-1,k,1)=\sum_{i=0}^{k-2} a(n+i,k,1), \qquad n \geqq 1.$$

THEOREM 4.8. *If* $1 \leqq n \leqq k-1$ *then* $a(n, k, 1) = 2^n - 1$.

Note that Result 4.7 and Theorem 4.8 provide a recurrence relation and sufficient initial conditions to easily compute $a(n, k, 1)$ for any reasonable values of n and k. Also note that in combination with Lemmas 4.2 and 4.3, the above two results have as an immediate corollary Result 2.2.

We have thus seen that Theorem 4.4 is the basis of straightforward proofs of all the results known previously on the computation of $a(n, k)$ for "small" k.

5. Proof of the main results.

In this section we prove the results given in § 4 above. We first consider Lemmas 4.1–4.3.

Proof of Lemma 4.1. Choose $\pi \in A(n, k)$, and suppose $i, j \in \{1, 2, \cdots, n\}$ satisfy $\bar{j} = \overline{i+1}$ (where, as always, the bars denote residue classes modulo n). By inspection:

$$X_k(\pi,i)-X_k(\pi,j)=\begin{cases}\{i\pi\} & \text{if } \overline{i\pi}\in\{\bar{i},\overline{i+1},\cdots,\overline{i+k-2}\},\\ \phi & \text{if } \overline{i\pi}=\overline{i+k-1},\end{cases}$$

and

$$X_k(\pi,j)-X_k(\pi,i)=\begin{cases}\{i+k-1\} & \text{if } \overline{i+k-1}\in\{\overline{(i+1)\pi},\cdots,\overline{(i+k-1)\pi}\},\\ \phi & \text{if } \overline{i\pi}=\overline{i+k-1}.\end{cases}$$

Hence $|X_k(\pi, i)| = |X_k(\pi, j)|$ and the result follows. □

Lemma 4.2 is immediate from the definition, and we also have the following proof.

Proof of Lemma 4.3. (i) Define the mapping ϕ_k from S_n into S_n by:

$$\phi_k(\pi) \text{ maps } i \text{ to } (n+1)-s\pi, \text{ where } s\in\{1,2,\cdots,n\} \text{ and } \bar{s}=\overline{-i-k+2}.$$

Then we claim that ϕ_k is a one-to-one mapping from $A(n, k, r)$ into $A(n, k, k - 1 - r)$. This will establish the result.

It is not difficult to see that ϕ_k permutes the elements of S_n, so we need only show that ϕ_k maps $A(n, k, r)$ into $A(n, k, k - 1 - r)$ to complete the proof of (i).

Choose $\pi \in A(n, k, r)$ and define $\pi^* = \phi_k(\pi)$. By definition, if $i \in \{1, 2, \cdots, n\}$ then $\overline{i\pi^*} = \overline{1 - s(i)\pi}$, where $s(i) \in \{1, 2, \cdots, n\}$ and $\overline{s(i)} = \overline{-i-k+2}$. Now since $\pi \in A(n, k)$, we know that

$$\overline{s(i)\pi} \in \{\overline{-i-k+2}, \overline{-i-k+3}, \cdots, \overline{-i+1}\},$$

and hence

$$\overline{i\pi^*} \in \{\overline{i+k-1}, \overline{i+k-2}, \cdots, \bar{i}\},$$

and so $\pi^* \in A(n, k)$. Now, by definition,

$$X_k(\pi^*, n-k+2) = \{j\pi^*, \overline{s(j)} \in \{\bar{n}, \overline{n-1}, \cdots, \overline{n-k+2}\} : \overline{s(j)\pi}$$
$$\in \{\overline{s(j)+k-1}, \overline{s(j)+k-2}, \cdots, \bar{1}\}\}$$

and since $\pi \in A(n, k, r)$, $|X_k(\pi, n - k + 2)| = r$, i.e.,

$$|\{j\pi, \bar{j} \in \{\overline{n-k+2}, \overline{n-k+3}, \cdots, \bar{n}\} : \overline{j\pi} \in \{\bar{j}, \overline{j+1}, \cdots, \bar{n}\}\}| = r,$$

and hence $|X_k(\pi^*, n - k + 2)| = k - 1 - r$, and (i) follows.

(ii) If $\pi \in A(n, k, 0)$, then it is straightforward to show that $\overline{i\pi} = \overline{i+k-1}$ for every i, and hence $|A(n, k, 0)| = 1$. The result follows from (i). □

In order to establish Theorem 4.4, it is necessary to prove a number of preliminary results. We first make some definitions.

If $0 \leqq r \leqq k - 1$ and $k \geqq 2$, then define $\mathbf{E}(k, r)$ to be the class of all $(k - 1)$-subsets E of $\{-k + 2, -k + 3, \cdots, k - 1\}$ satisfying the property that E contains precisely r elements of $\{-k + 2, -k + 3, \cdots, 0\}$.

If $E \in \mathbf{E}(k, r)$ then define $U_k(E)$ to be the set:

$$\{(c_1, c_2, \cdots, c_{k-1}) : \{c_1, c_2, \cdots, c_{k-1}\} = E, c_i \in \{i-k+1, i-k+2, \cdots, i\}\}.$$

In addition let $u_k(E) = |U_k(E)|$.

Lastly, for any set of integers $E = \{e_1, e_2, \cdots, e_s\}$, say, let $\bar{E} = \{\overline{e_1}, \overline{e_2}, \cdots, \overline{e_s}\}$, where, as always, we are working modulo n.

As an immediate result we have the following lemma.

LEMMA 5.1.

$$|\mathbf{E}(k, r)| = \binom{k-1}{r}^2, \qquad 0 \leqq r \leqq k-1, \quad k \geqq 2.$$

We may now state the following important result, which justifies the definition of $U_k(E)$.

LEMMA 5.2. *Suppose* $0 \leqq r \leqq k - 1$ *and* $2 \leqq k \leqq n$. *Then* $(c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$ *for some* $E = \{c_1, c_2, \cdots, c_{k-1}\}$ *satisfying* $E \in \mathbf{E}(k, r)$ *and* $|\bar{E}| = k - 1$, *if and only if there exists* $\pi \in A(n, k, r)$ *satisfying*:

$$j\pi = \begin{cases} c_i & \text{if } j\pi < j \\ c_i + n & \text{if } j\pi \geqq j \end{cases} \quad \text{where } j = n-k+1+i, \quad i \in \{1, 2, \cdots, k-1\}.$$

Proof. First suppose $\pi \in A(n, k, r)$, and let $(c_1, c_2, \cdots, c_{k-1})$ be as in the statement of the lemma. Then if $E = \{c_1, c_2, \cdots, c_{k-1}\}$, we must show the following:

(i) $|\bar{E}| = k - 1$,

(ii) $E \in \mathbf{E}(k, r)$,

(iii) $(c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$.

Item (i) follows since $\overline{c_i} = \overline{(n-k+1+i)\pi}$ for every $i \in \{1, 2, \cdots, k-1\}$ and $\overline{c_i} = \overline{c_j}$ if and only if $i = j$ (since $\pi \in S_n$).

Next, since $\pi \in A(n, k)$, we know that

$$\overline{(n-k+1+i)\pi} \in \{\overline{n-k+1+i}, \overline{n-k+2+i}, \cdots, \overline{n+i}\}.$$

Hence if $(n-k+1+i)\pi \geqq (n-k+1+i)$, then

$$(n-k+1+i)\pi \in \{n-k+1+i, n-k+2+i, \cdots, n\},$$

i.e.,

$$c_i = (n-k+1+i)\pi - n \in \{-k+1+i, -k+2+i, \cdots, 0\}.$$

Similarly, if $(n-k+1+i)\pi < (n-k+1+i)$, then

$$c_i = (n-k+1+i)\pi \in \{1, 2, \cdots, i\}.$$

We have thus shown that $E \in \mathbf{E}(k, s)$, where

$$s = |\{j \in \{n-k+2, n-k+3, \cdots, n\} : j\pi \geqq j\}|,$$

and moreover that $c_i \in \{i-k+1, i-k+2, \cdots, i\}$ for every i, and so we have shown (iii).

Now $\pi \in A(n, k, r)$, and hence $|X_k(\pi, n-k+2)| = r$, i.e.,

$$|\{j\pi, j \in \{n-k+2, n-k+3, \cdots, n\} : \overline{j\pi} \in \{\bar{j}, \overline{j+1}, \cdots, \bar{n}\}\}| = r.$$

But $\overline{j\pi} \in \{\bar{j}, \overline{j+1}, \cdots, \bar{n}\}$ iff $j\pi \in \{j, j+1, \cdots, n\}$ iff $j\pi \geqq j$. Hence $s = r$ and (ii) follows.

Now suppose $(c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$, where $\{c_1, c_2, \cdots, c_{k-1}\} = E$, and $E \in \mathbf{E}(k, r)$ has the property: $|\bar{E}| = k-1$.

If $D = \{\bar{1}, \bar{2}, \cdots, \bar{n}\} - \bar{E}$, then $|D| = n-k+1$. Hence let $\{d_1, d_2, \cdots, d_{n-k+1}\}$ be the set satisfying the following three properties:

(i) $D = \{\overline{d_1}, \overline{d_2}, \cdots, \overline{d_{n-k+1}}\}$,

(ii) $d_i \in \{1, 2, \cdots, n\}$ for every $i \in \{1, 2, \cdots, n-k+1\}$, and

(iii) $d_i < d_{i+1}$ for every $i \in \{1, 2, \cdots, n-k\}$.

Note that by (ii) and (iii) it is immediate that

(1) $$i \leqq d_i \leqq i+k-1 \quad \text{for every } i \in \{1, 2, \cdots, n-k+1\}.$$

Define $\pi \in S_n$ as follows. Let:

$$j\pi = \begin{cases} d_j & \text{if } 1 \leqq j \leqq n-k+1, \\ c_i & \text{if } n-k+2 \leqq j \leqq n \text{ and } c_i > 0 \quad \text{where } i = j-n+k-1, \\ c_i + n & \text{if } n-k+2 \leqq j \leqq n \text{ and } c_i \leqq 0 \quad \text{where } i = j-n+k-1. \end{cases}$$

It is clear that π is well defined, since, by definition,

$$\{\overline{c_1}, \overline{c_2}, \cdots, \overline{c_{k-1}}, \overline{d_1}, \overline{d_2}, \cdots, \overline{d_{n-k+1}}\} = \{\bar{1}, \bar{2}, \cdots, \bar{n}\}.$$

Again by definition, $d_i \in \{1, 2, \cdots, n\}$ for every i. Finally if $c_i > 0$, then

$$c_i \in \{1, 2, \cdots, k-1\},$$

and if $c_i \leqq 0$, then $c_i \in \{-k+2, -k+3, \cdots, 0\}$, and hence

$$c_i + n \in \{n-k+2, n-k+3, \cdots, n\}.$$

Now suppose $j = n-k+1+i$, $i \in \{1, 2, \cdots, k-1\}$. Then

$$j\pi = \begin{cases} c_i & \text{if } c_i > 0 \\ c_i + n & \text{if } c_i \leqq 0 \end{cases} \quad \text{and } c_i \in \{i-k+1, i-k+2, \cdots, i\} \text{ for all } i.$$

Hence if $j\pi = c_i$ then $c_i > 0$, and so $c_i \in \{1, 2, \cdots, i\}$, i.e., $j\pi \leqq i$. Now

$$j = n - k + 1 + i > i,$$

i.e., $j\pi < j$. Similarly, if $j\pi = c_i + n$, then we have $c_i \leqq 0$, and so

$$c_i \in \{i - k + 1, i - k + 2, \cdots, 0\},$$

i.e., $j\pi \geqq n - k + 1 + i = j$. Hence

$$j\pi = \begin{cases} c_i & \text{if } j\pi < j, \\ c_i + n & \text{if } j\pi \geqq j. \end{cases}$$

We now only need show that $\pi \in A(n, k, r)$. By (1), $j \leqq d_j \leqq j + k - 1$ and hence $\overline{j\pi} \in \{\bar{j}, \overline{j+1}, \cdots, \overline{j+k-1}\}$ for every $j \in \{1, 2, \cdots, n - k + 1\}$. Also, if $n - k + 2 \leqq j \leqq n$, then $\overline{j\pi} = \overline{c_i} \in \{\overline{i-k+1}, \overline{i-k+2}, \cdots, \bar{i}\}$, where $j = n - k + 1 + i$. This implies that $\overline{j\pi} \in \{\bar{j}, \overline{j+1}, \cdots\ \overline{j+k-1}\}$ for every $j \in \{n - k + 2, n - k + 3, \cdots, n\}$, and so $\pi \in A(n, k)$.

Finally, by definition,

$$\begin{aligned} X_k(\pi, n-k+2) &= \{j\pi, j \in \{n-k+2, n-k+3, \cdots, n\} : \overline{j\pi} \in \{\bar{j}, \overline{j+1}, \cdots, \bar{n}\}\} \\ &= \{j\pi, j \in \{n-k+2, n-k+3, \cdots, n\} : j\pi \geqq j\}. \end{aligned}$$

Hence, by the above arguments,

$$|X_k(\pi, n-k+2)| = |\{c_i \in E : c_i \leqq 0\}| = r, \text{ since } E \in \mathbf{E}(k, r).$$

The result follows. □

The above result gives us a means of classifying the "endings" of permutations in $A(n, k, r)$, where the ending of a permutation π is the $(k - 1)$-tuple $((n - k + 2)\pi, (n - k + 3)\pi, \cdots, n\pi)$. The next result gives us a way of enumerating the number of "starts" for each possible ending.

LEMMA 5.3. *Suppose* $0 \leqq r \leqq k - 1$, $2 \leqq k \leqq n$, *and let* $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1})$ *and* $\mathbf{d} = (d_1, d_2, \cdots, d_{k-1})$ *be elements of* $U_k(E)$ *for some* $E \in \mathbf{E}(k, r)$. *If* $P(\mathbf{c})$ *is the set of permutations* $\pi \in A(n, k, r)$ *satisfying*

$$j\pi = \begin{cases} c_i & \text{if } j\pi < j, \\ c_i + n & \text{if } j\pi \geqq j, \end{cases} \quad j = n - k + 1 + i, \quad i \in \{1, 2, \cdots, k-1\},$$

and $P(\mathbf{d})$ *is the set of permutations* $\pi^* \in A(n, k, r)$ *satisfying*

$$j\pi^* = \begin{cases} d_i & \text{if } j\pi^* < j, \\ d_i + n & \text{if } j\pi^* \geqq j, \end{cases} \quad j = n - k + 1 + i, \quad i \in \{1, 2, \cdots, k-1\},$$

then $|P(\mathbf{c})| = |P(\mathbf{d})|$.

Proof. We define ϕ which maps $P(\mathbf{c})$ into $P(\mathbf{d})$ by:

$$i\phi(\pi) = \begin{cases} i\pi & \text{if } 1 \leqq i \leqq n-k+1 \\ i\pi^* & \text{if } n-k+2 \leqq i \leqq n \end{cases} \quad \text{where } \pi^* \text{ is any element of } P(\mathbf{d}).$$

We now show why ϕ is well defined. First suppose π^* and $\pi^{*\prime}$ are two elements of $P(\mathbf{d})$, and then, by definition, $i\pi^* = i\pi^{*\prime}$ for every $i \in \{n - k + 2, n - k + 3, \cdots, n\}$. Second, $\phi(\pi) \in S_n$, since if $\pi \in A(n, k, r)$ satisfies

$$j\pi = \begin{cases} c_i & \text{if } j\pi < j, \\ c_i + n & \text{if } j\pi \geqq j, \end{cases} \quad j = n - k + 1 + i, \quad i \in \{1, 2, \cdots, k-1\},$$

$\pi^* \in A(n, k, r)$ satisfies

$$j\pi^* = \begin{cases} d_i & \text{if } j\pi < j, \\ d_i + n & \text{if } j\pi \geqq j, \end{cases} \quad j = n-k+1+i, \quad i \in \{1, 2, \cdots, k-1\}$$

and $(c_1, c_2, \cdots, c_{k-1})$, $(d_1, d_2, \cdots, d_{k-1}) \in U_k(E)$, then it is clear that

$$\{j\pi : n-k+2 \leqq j \leqq n\} = \{j\pi^* : n-k+2 \leqq j \leqq n\}.$$

Third, it is straightforward to see that $\phi(\pi)$ is an element of $A(n, k)$ since $\pi, \pi^* \in A(n, k)$. Fourth, since $X_k(\phi(\pi), n-k+2) = X_k(\pi^*, n-k+2)$, and since $\pi^* \in A(n, k, r)$, it is clear that $\phi(\pi) \in A(n, k, r)$. Finally, by definition it is clear that $\phi(\pi) \in P(\mathbf{d})$. We have thus shown that ϕ is well defined.

To conclude the proof we show that ϕ is one to one. Suppose that $\phi(\pi) = \phi(\pi')$, where $\pi, \pi' \in A(n, k, r)$ and where $\pi, \pi' \in P(\mathbf{c})$. Then $i\pi = i\pi'$ for every i satisfying $1 \leqq i \leqq n-k+1$. But since $\pi, \pi' \in P(\mathbf{c})$ we know that $i\pi = i\pi'$ for every

$$i \in \{n-k+2, n-k+3, \cdots, n\}.$$

Hence, $\pi = \pi'$ and the result follows. □

Because of Lemma 5.3 we can make the following definition, the relevance of which is apparent in the next result. If $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$, and $P(\mathbf{c})$ is as in the statement of Lemma 5.3, then let $v_{n,k}(E) = |P(\mathbf{c})|$. $v_{n,k}(E)$ is well defined precisely because of Lemma 5.3. We can now state the following important result.

THEOREM 5.4. *If* $0 \leqq r \leqq k-1$ *and* $2 \leqq k \leqq n$, *then*

$$a(n,k,r) = \sum_{E \in \mathbf{E}(k,r)} u_k(E) v_{n,k}(E).$$

Proof. By definition,

$$\begin{aligned} a(n,k,r) &= |A(n,k,r)| \\ &= \sum_* \left(\sum_{\mathbf{c} \in U_k(E)} |P(\mathbf{c})| \right) \quad \text{(by Lemma 5.2)} \\ &\left(\text{where } \sum_* \text{ denotes the sum over all } E \in \mathbf{E}(k,r) \text{ satisfying } |\bar{E}| = k-1 \right) \\ &= \sum_* u_k(E) v_{n,k}(E) \quad \text{(by Lemma 5.3)} \\ &= \sum_{E \in \mathbf{E}(k,r)} u_k(E) v_{n,k}(E) \end{aligned}$$

since if $|\bar{E}| < k-1$, then it is clear that $v_{n,k}(E) = 0$. □

We have thus transformed the problem of evaluating $a(n, k, r)$ into the problem of evaluating $v_{n,k}(E)$ and $u_k(E)$ for every $E \in \mathbf{E}(k, r)$. In the next two results we show how these values may be computed.

THEOREM 5.5. *Suppose* $0 \leqq r \leqq k-1$ *and* $2 \leqq k \leqq n$, *and let* $E \in \mathbf{E}(k, r)$. *Then*

(i) *If* $n = k$ *then*

$$v_{n,k}(E) = \begin{cases} 1 & \textit{if } |\bar{E}| = k-1, \\ 0 & \textit{if } |\bar{E}| < k-1; \end{cases}$$

(ii) *$v_{n+1,k}(E) = \sum_* v_{n,k}(F)$, where $\sum_*$ represents the sum over all $F \in \mathbf{E}(k, r)$ which contain the set E^* which is defined to be the union of $\{i : i \in E, i > 0\}$ and $\{i + 1 : i \in E, i < 0\}$.*

Proof. First note that $U_k(E)$ is nonempty for any $E \in \mathbf{E}(k, r)$, since an element of $U_k(E)$ can always be produced by assemblying the elements of E in ascending order.

(i) Suppose $n = k$. First let $|\bar{E}| < k - 1$, and then, using the notation of Lemma 5.3, suppose that $\pi \in A(n, k, r)$ is an element of $P(\mathbf{c})$ for some $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$. Then, since $|\bar{E}| < k - 1$, there exists a pair c_i, c_j $(i \neq j)$ with $\bar{c}_i = \bar{c}_j$. Hence $\overline{(i+1)\pi} = \overline{(j+1)\pi}$, i.e., $(i + 1)\pi = (j + 1)\pi$, which is a contradiction since π is a permutation. Hence $v_{k,k}(E) = 0$ if $|\bar{E}| < k - 1$.

Now suppose $|\bar{E}| = k - 1$, and choose a $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$. If $\pi \in P(\mathbf{c})$ (π exists by Lemma 5.2) then, by definition, $\{\overline{2\pi}, \overline{3\pi}, \cdots, \overline{k\pi}\} = \bar{E}$, and hence $\{\overline{1\pi}\} = \{\bar{1}, \bar{2}, \cdots, \bar{k}\} - \bar{E}$. So 1π is fixed by the choice of E, and, by definition, $2\pi, 3\pi, \cdots, k\pi$ are also fixed since $\pi \in P(\mathbf{c})$. Thus π is uniquely defined, and so $v_{k,k}(E) = 1$.

(ii) First note that the elements of E are all distinct modulo $n + 1$ if and only if the elements of E^* are all distinct modulo n. Hence we assume that both these statements are true, since otherwise both sides of the equation are zero by (i) above.

First, choose $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$. If $0 \in E$, then let h satisfy $c_h = 0$, and, if $0 \notin E$, then set $h = 0$. Then, by definition, if $\pi \in A(n + 1, k, r)$ satisfies $\pi \in P(\mathbf{c})$ we have

$$(n-k+2+h)\pi = n+1.$$

Next, suppose that $F = \{d_1, d_2, \cdots, d_{k-1}\} \in \mathbf{E}(k, r)$ contains E^*.

Then, if $0 \notin E$ we have $F = E^*$, and we let

$$d_i = \begin{cases} c_i + 1 & \text{if } c_i < 0, \\ c_i & \text{if } c_i > 0, \end{cases} \qquad 1 \leqq i \leqq k-1.$$

Note that if $0 \notin E$ then it is clear that $E^* \in \mathbf{E}(k, r)$.

If $0 \in E$, then we let $d_1, d_2, \cdots, d_{k-1}$ be defined as follows:

d_1 is the element of $\{-k + 2, -k + 3, \cdots, 0\} - E^*$ that is contained in F.

$$d_i = \begin{cases} c_{i-1} + 1 & \text{if } c_{i-1} < 0, \\ c_{i-1} & \text{if } c_{i-1} > 0, \end{cases} \qquad 2 \leqq i \leqq h,$$

$$d_i = \begin{cases} c_i + 1 & \text{if } c_i < 0, \\ c_i & \text{if } c_i > 0, \end{cases} \qquad h+1 \leqq i \leqq k-1.$$

Now let $\mathbf{d}_F = (d_1, d_2, \cdots, d_{k-1})$ and we now show that $\mathbf{d}_F \in U_k(F)$. To do this we need only show that $d_i \in \{i - k + 1, i - k + 2, \cdots, i\}$ for every $i \in \{1, 2, \cdots, k - 1\}$. First, suppose that $i \leqq h$:

If $i = 1$, then $d_1 \in \{-k + 2, -k + 3, \cdots, 0\} - E^*$, i.e.,

$$d_1 \in \{-k+2, -k+3, \cdots, 1\}.$$

If $i > 1$, then we have

$$d_i = \begin{cases} c_{i-1} + 1 & \text{if } c_{i-1} < 0, \\ c_{i-1} & \text{if } c_{i-1} > 0. \end{cases}$$

If $c_{i-1} < 0$ then $d_i = c_{i-1} + 1 \in \{i - k + 1, i - k + 2, \cdots, i\}$, since $\mathbf{c} \in U_k(E)$.
If $c_{i-1} > 0$ then $d_i = c_{i-1} \in \{1, 2, \cdots, i - 1\}$, i.e.,

$$d_i \in \{i-k+2, i-k+3, \cdots, i\},$$

since $\mathbf{c} \in U_k(E)$ and $c_{i-1} > 0$. Second, suppose that $i > h$:
Then we have

$$d_i = \begin{cases} c_i + 1 & \text{if } c_i < 0, \\ c_i & \text{if } c_i > 0. \end{cases}$$

If $c_i < 0$ then $d_i = c_i + 1 \in \{i - k + 2, i - k + 3, \cdots, 0\}$, since $\mathbf{c} \in U_k(E)$.
If $c_i > 0$ then $d_i = c_i \in \{1, 2, \cdots, i\}$ (since $\mathbf{c} \in U_k(E)$), i.e.,

$$d_i \in \{i-k+1, i-k+2, \cdots, i\}.$$

Hence $\mathbf{d}_F \in U_k(F)$.

Now, using the same notation as before, let $P(\mathbf{c})$ and $P(\mathbf{d}_F)$ be sets of permutations from $A(n + 1, k)$ and $A(n, k)$ respectively, defined as in the statement of Lemma 5.3. We will show that

$$|P(\mathbf{c})| = \sum_{*} |P(\mathbf{d}_F)|,$$

where, as in the statement of the theorem, $\sum_*$ represents the sum over all $F \in \mathbf{E}(k, r)$ which contain E^*. This will establish the result. We actually prove this claim by exhibiting a one-to-one correspondence ϕ between $P(\mathbf{c})$ and the union of the sets $P(\mathbf{d}_F)$, which are clearly all disjoint. We define ϕ as follows:

Suppose $\pi \in A(n + 1, k, r)$ is contained in $P(\mathbf{c})$, i.e., suppose that

$$j\pi = \begin{cases} c_i & \text{if } j\pi < j, \\ c_i + n + 1 & \text{if } j\pi \geqq j, \end{cases} \qquad j = n-k+2+i, \quad i \in \{1, 2, \cdots, k-1\}.$$

Then define $\pi^* = \phi(\pi)$ by

$$i\pi^* = \begin{cases} i\pi & \text{if } 1 \leqq i \leqq n-k+1+h, \\ (i+1)\pi & \text{if } n-k+2+h \leqq i \leqq n. \end{cases}$$

We now show that π^* is an element of $P(\mathbf{d}_F)$, where $F \in \mathbf{E}(k, r)$ contains E^*, and hence show that ϕ is well defined.

First, note that $\pi^* \in S_n$ since $\pi \in S_{n+1}$ and h is chosen so that $(n - k + 2 + h)\pi = n + 1$.

Second, observe that $\pi^* \in A(n, k)$. We show this as follows:

If $1 \leqq i \leqq n - k + 1$ then, since $\pi \in A(n + 1, k)$, we have

$$i\pi^* = i\pi \in \{i, i+1, \cdots, i+k-1\}.$$

If $n - k + 2 \leqq i \leqq n - k + 1 + h$ (which only applies if $h > 0$) then, since $\pi \in A(n + 1, k)$, $i\pi \in \{i, i + 1, \cdots, n, 1, 2, \cdots, i - n + k - 2\}$; note that $i\pi \neq n + 1$ since $i \neq n - k + 2 + h$. Hence $\overline{i\pi^*} \in \{\bar{i}, \overline{i+1}, \cdots, \bar{n}, \bar{1}, \bar{2}, \cdots, \overline{i-n+k-2}\}$, i.e., $\overline{i\pi^*} \in \{\bar{i}, \overline{i+1}, \cdots, \overline{i+k-2}\}$.

If $n - k + 2 + h \leqq i \leqq n$ then, since $\pi \in A(n + 1, k)$, we have

$$i\pi^* = (i+1)\pi \in \{i+1, i+2, \cdots, n, 1, 2, \cdots, i-n+k-1\};$$

note that $i\pi^* \neq n + 1$ since $i + 1 \neq n - k + 2 + h$. Hence

$$\overline{i\pi^*} \in \{\overline{i+1}, \overline{i+2}, \cdots, \overline{i+k-1}\}.$$

Thus $\pi^* \in A(n, k)$.

Third, note that $\pi^* \in A(n, k, r)$. This can be demonstrated by considering $X_k(\pi^*, n - k + 2)$. By definition,

$$\begin{aligned} &X_k(\pi^*, n-k+2) \\ &\quad = \{j\pi^*, j \in \{n-k+2, n-k+3, \cdots, n\} : j\pi^* \in \{j, j+1, \cdots, n\}\} \\ &\quad = \text{the union of } \{j\pi, j \in \{n-k+2, n-k+3, \cdots, \\ &\qquad\qquad\qquad n-k+1+h\} : j\pi \in \{j, j+1, \cdots, n\}\} \end{aligned}$$

and

$$\{j\pi, j \in \{n-k+3+h, n-k+4+h, \cdots, n+1\} : j\pi \in \{j\pi \in \{j-1, j, \cdots, n\}\}\}.$$

Now since $\pi \in A(n + 1, k)$, where $n + 1 > n \geqq k$, we know that $j\pi \neq j - 1$ for any j, and hence

$$\begin{aligned} X_k(\pi^*, n-k+2) = \{j\pi, j \in \{n-k+2, n-k+3, \cdots, n\}, \\ j \neq n-k+2+h : j\pi \in \{j, j+1, \cdots, n\}\}. \end{aligned}$$

Also note that $(n - k + 2 + h)\pi = n + 1$ and hence $X_k(\pi^*, n - k + 2) = X_k(\pi, n - k + 2)$ and hence $\pi^* \in A(n, k, r)$.

Fourth, we let

$$F = \begin{cases} E^* & \text{if } h = 0, \\ \text{the union of } E^* \text{ and } \{(n-k+2)\pi - n\} & \text{if } h > 0. \end{cases}$$

Then F contains E^* by definition. We claim that $F \in \mathbf{E}(k, r)$, and, defining $\mathbf{d}_F$ as above, we also claim that $\pi^* \in P(\mathbf{d}_F)$.

We first show that $F \in \mathbf{E}(k, r)$.

If $h = 0$ then $0 \notin E$ and hence $F = E^* \in \mathbf{E}(k, r)$.

If $h > 0$ then $0 \in E$ and hence E^* contains $r - 1$ elements of

$$\{-k+2, -k+3, \cdots, 0\}$$

and $k - 1 - r$ elements of $\{1, 2, \cdots, k - 1\}$. Now since

$$\pi \in A(n+1, k), (n-k+2)\pi \in \{n-k+2, n-k+3, \cdots, n+1\},$$

and hence $(n - k + 2)\pi - n \in \{-k + 2, -k + 3, \cdots, 0, 1\}$. Now since $h > 0$, $(n - k + 2)\pi \neq n + 1$, i.e., $(n - k + 2)\pi - n \neq 1$. Hence

$$(n-k+2)\pi - n \in \{-k+2, -k+3, \cdots, 0\},$$

and so to show that $F \in \mathbf{E}(k, r)$ we need only show that $(n - k + 2)\pi - n \notin E^*$. But since $\pi \in A(n + 1, k)$, $(n - k + 2)\pi - (n + 1) \notin E$ and the result follows.

To see that $\pi^* \in P(\mathbf{d}_F)$ we need only examine the values of $j\pi^*$, where $j = n - k + 1 + i$ and $i \in \{1, 2, \cdots, k - 1\}$. Choose such a j.

If $i > h$ then

$$\begin{aligned} j\pi^* &= (j+1)\pi \\ &= \begin{cases} c_i & \text{if } j\pi < j \\ c_i + n + 1 & \text{if } j\pi \geqq j \end{cases} \quad \text{since } \pi \in P(\mathbf{c}) \end{aligned}$$

$$= \begin{cases} d_i & \text{if } j\pi < j \\ d_i + n & \text{if } j\pi \geqq j \end{cases} \quad \text{since, given } i > h,$$

$$d_i = \begin{cases} c_i & \text{if } c_i > 0 \\ c_i + 1 & \text{if } c_i < 0 \end{cases} \quad \text{and } c_i < 0 \text{ iff } j\pi \geqq j.$$

If $i \leqq h$ then we have two cases to consider: $i = 1$ and $i > 1$.

If $i = 1$ then

$$j\pi^* = (n-k+2)\pi^* = (n-k+2)\pi \in \{n-k+2, n-k+3, \cdots, n\}, \text{ i.e., } j\pi \geqq j.$$

Now, by the above, $(n - k + 2)\pi - n \in F$, $(n - k + 2)\pi - n \in \{-k + 2, -k + 3, \cdots, 0\}$ and $(n - k + 2)\pi - n \notin E^*$. Hence $d_1 = (n - k + 2)\pi - n$, i.e., $(n - k + 2)\pi^* = d_1 + n$ and $(n - k + 2)\pi \geqq n - k + 2$.

If $2 \leqq i \leqq h$ then

$$j\pi^* = j\pi = \begin{cases} c_{i-1} & \text{if } j\pi < j, \\ c_{i-1} + n + 1 & \text{if } j\pi \geqq j, \end{cases}$$

$$= \begin{cases} d_i & \text{if } j\pi < j, \\ d_i + n & \text{if } j\pi \geqq j, \end{cases}$$

since, given $i \leqq h$,

$$d_i = \begin{cases} c_{i-1} & \text{if } c_{i-1} > 0 \\ c_{i-1} + 1 & \text{if } c_{i-1} < 0 \end{cases} \quad \text{and } c_{i-1} < 0 \text{ iff } j\pi \geqq j.$$

Hence $\pi^* \in P(\mathbf{d}_F)$ and we have shown that ϕ is well defined.

To complete the proof, we need to show that ϕ is one to one and onto.

First, suppose that $\pi_1, \pi_2 \in P(\mathbf{c})$ satisfy $\phi(\pi_1) = \phi(\pi_2)$. Then, by definition of ϕ, $i\pi_1 = i\pi_2$ for every $i \in \{1, 2, \cdots, n + 1\}$ except for $i = n - k + 2 + h$. However, since π_1 and π_2 are permutations, they cannot disagree in exactly one position and hence $\pi_1 = \pi_2$ and thus we have shown that ϕ is one to one.

We now show that ϕ is onto, and hence complete the proof. Suppose that $\pi^* \in A(n, k, r)$ is contained in $P(\mathbf{d}_F)$, where $F \in \mathbf{E}(k, r)$ contains E^*.

Then let $\pi \in S_{n+1}$ satisfy

$$i\pi = \begin{cases} i\pi^* & \text{if } 1 \leqq i \leqq n-k+1+h, \\ n+1 & \text{if } i = n-k+2+h, \\ (i-1)\pi^* & \text{if } n-k+3+h \leqq i \leqq n+1. \end{cases}$$

Note that π is clearly in S_{n+1} since $\pi^* \in S_n$.

It is now straightforward to verify that $\pi \in A(n + 1, k, r)$, and moreover that $\pi \in P(\mathbf{c})$ and $\phi(\pi) = \pi^*$. This establishes that ϕ is onto and the result follows. □

THEOREM 5.6. *Suppose* $0 \leqq r \leqq k - 1$, $2 \leqq k$ *and* $n = 2k - 2$, *and let* $E \in \mathbf{E}(k, r)$. *Then* $u_k(E) = v_{n,k}(F)$, *where* $F \in \mathbf{E}(k, r)$ *is defined by*

$$F = \{i \in \{-k+2, -k+3, \cdots, k-1\} : \bar{i} = \overline{j+k-1},$$
$$j \in \{-k+2, -k+3, \cdots, k-1\} - E\}.$$

Proof. We first show that F as defined in the statement of the theorem is always in $\mathbf{E}(k, r)$. By definition, $F \in \mathbf{E}(k, s)$, where

$$s = |\{i \in F : i \in \{-k+2, -k+3, \cdots, 0\}\}|,$$

and so we need only show that $r = s$.

Suppose $i \in \{-k+2, -k+3, \cdots, 0\}$. Then, by definition, $i \in F$ if and only if $\overline{i-k+1} \notin \bar{E}$, since $n = 2k-2$ (note that bars denote residue classes modulo $n = 2k-2$). Hence, again by definition,

$$\begin{aligned} s &= k-1-|\{i \notin F : i \in \{-k+2, -k+3, \cdots, 0\}\}| \\ &= k-1-|\{\bar{i} \in \bar{E} : \bar{i} \in \{\bar{1}, \bar{2}, \cdots, \overline{k-1}\}\}| \\ &= r \quad (\text{since } E \in \mathbf{E}(k, r)) \quad \text{and thus } F \in \mathbf{E}(k, r). \end{aligned}$$

Now choose an element $\mathbf{d} = (d_1, d_2, \cdots, d_{k-1})$ from $U_k(F)$. We must show (using the above notation) that $|U_k(E)| = |P(\mathbf{d})|$, and we will then have completed the proof. To do this we define ϕ which maps $U_k(E)$ into $P(\mathbf{d})$, as follows:

Suppose $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1}) \in U_k(E)$. Then $\pi = \phi(\mathbf{c})$ satisfies

(i)

$$\overline{i\pi} = \begin{cases} \overline{c_i+k-1} & \text{if } 1 \leqq i \leqq k-1, \\ \bar{d}_{i-k+1} & \text{if } k \leqq i \leqq 2k-2, \end{cases}$$

(ii) $i\pi \in \{1, 2, \cdots, n\}$, $1 \leqq i \leqq 2k-2$.

We must first show that ϕ is well defined, i.e., that $\pi \in P(\mathbf{d})$. We first show that $\pi \in S_n$. By definition, π maps $\{1, 2, \cdots, n\}$ into $\{1, 2, \cdots, n\}$, and hence we need only show that π is one to one. Also, since $\bar{c}_i \neq \bar{c}_j$ and $\bar{d}_i \neq \bar{d}_j$ $(i \neq j)$, we need only show that $\overline{c_i+k-1} \neq \bar{d}_j$ for any $i, j \in \{1, 2, \cdots, k-1\}$.

Suppose $\overline{c_i+k-1} = \bar{d}_j$; then, by definition of F, $\bar{d}_j = \overline{s+k-1}$, where $s \in \{-k+2, -k+3, \cdots, k-1\} - E$. Hence $\bar{c}_i = \bar{s}$, where $s \notin E$ and $c_i \in E$. This gives us the required contradiction, and hence $\pi \in S_n$.

We next show that $\pi \in A(n, k)$. If $1 \leqq i \leqq k-1$, then

$$\overline{i\pi} = \overline{c_i+k-1} \in \{\bar{i}, \overline{i+1}, \cdots, \overline{i+k-1}\} \quad (\text{since } \mathbf{c} \in U_k(E)).$$

If $k \leqq i \leqq 2k-2$, then $\overline{i\pi} = \bar{d}_{i-k+1} \in \{\overline{i-2k+2}, \overline{i-2k+3}, \cdots, \overline{i-k+1}\}$ (since $\mathbf{d} \in U_k(F)) = \{\bar{i}, \overline{i+1}, \cdots, \overline{i+k-1}\}$ (since $n = 2k-2$). Hence $\pi \in A(n, k)$.

Next observe that, by Lemma 5.2, since $F \in \mathbf{E}(k, r)$ and $\mathbf{d} \in U_k(F)$ there exists a $\pi^* \in A(n, k, r)$ satisfying $\overline{i\pi^*} = \overline{i\pi}$ for every $i \in \{k, k+1, \cdots, 2k-2\}$. Hence $X_k(\pi, k) = X_k(\pi^*, k)$, and so $\pi \in A(n, k, r)$. Finally, note that $\pi \in P(\mathbf{d})$ by definition, and so ϕ is well defined.

By definition it is clear that ϕ is one to one, and so to complete the proof we need only show that ϕ is onto. Suppose $\pi \in P(\mathbf{d})$. We must show that if $\mathbf{c} = (c_1, c_2, \cdots, c_{k-1})$ satisfies

(i) $\bar{c}_i = \overline{i\pi-k+1}$, and

(ii) $c_i \in \{-k+2, -k+3, \cdots, k-1\}$ for every i, then $\mathbf{c} \in U_k(E)$.

Since $\pi \in A(n, k, r)$, $\overline{i\pi} \in \{\bar{i}, \overline{i+1}, \cdots, \overline{i+k-1}\}$, and hence $\bar{c}_i = \overline{i\pi-k+1} \in \{\overline{i-k+1}, \overline{i-k+2}, \cdots, \bar{i}\}$. Thus we need only show that $\{c_1, c_2, \cdots, c_{k-1}\} = E$. By definition, $\pi \in P(\mathbf{d})$, and hence

$$\{\overline{k\pi}, \overline{(k+1)\pi}, \cdots, \overline{(2k-2)\pi}\} = \bar{F} = \{\bar{i} : 1 \leqq i \leqq n, \bar{i} = \overline{j+k-1},$$
$$j \in \{-k+2, -k+3, \cdots, k-1\} - E\}.$$

Thus, since $n = 2k - 2$,

$$\{\overline{k\pi+k-1},\overline{(k+1)\pi+k-1},\cdots,\overline{(2k-2)\pi+k-1}\}=\{\bar{1},\bar{2},\cdots,\bar{n}\}-\bar{E}.$$

Finally, since $\pi \in S_n$, $\{\bar{c}_1, \bar{c}_2, \cdots, \bar{c}_{k-1}\} = \bar{E}$, and since $n = 2k - 2$,

$$\{c_1,c_2,\cdots,c_{k-1}\}=E. \qquad \square$$

Theorems 5.5 and 5.6 now enable us to prove the main result of this paper, namely Theorem 4.4.

Proof of Theorem 4.4. First suppose $0 \leqq r \leqq k - 1$ and $2 \leqq k$. Then, as before we let

$$t=\binom{k-1}{r}=\binom{k-1}{k-1-r}.$$

As in the definition of $H(k, r)$ preceding the statement of Theorem 4.4, label the t distinct r-subsets of $\{-k + 2, -k + 3, \cdots, 0\} : (R_1, R_2, \cdots, R_t)$, and let

$$R_i^* = \{j+1 : j \in R_i-\{0\}\}$$

for every i. Then $H(k, r) = (h_{ij})$ satisfies

$$h_{ij}=\begin{cases}1 & \text{if } R_i^* \text{ is a subset of } R_j,\\ 0 & \text{otherwise.}\end{cases}$$

We need to show that $a(n, k, r) = \text{Trace}\,(H(k, r)^n)$ for every $n \geqq k$.

For every $i \in \{1, 2, \cdots, t\}$ define

$$C_i=\{j+k-1 : j \in \{-k+2,-k+3,\cdots,0\}-R_i\}.$$

Then C_i is a $(k - 1- r)$-subset of $\{1, 2, \cdots, k - 1\}$ for every i, and $(C_1, C_2, \cdots, C_t)$ forms a labeling of all such subsets. Now let

$$X_{ij}=\{s : s \in R_i \text{ or } s \in C_j\},$$

i.e., X_{ij} is the union of R_i and C_j. Then it is clear that

$$\mathrm{E}(k,r)=\{X_{ij} : 1 \leqq i \leqq t, 1 \leqq j \leqq t\}.$$

Next, for every $n \geqq k$, define the t by t matrix $W(n) = (w(n)_{ij})$ by $w(n)_{ij} = v_{n,k}(X_{ij})$.

We first consider $W(k)$. By Theorem 5.5 (i),

$$w(k)_{ij}=v_{k,k}(X_{ij})=\begin{cases}1 & \text{if } \bar{R}_i \text{ and } \bar{C}_j \text{ are disjoint,}\\ 0 & \text{otherwise,}\end{cases}$$

where the bars denote residue classes modulo k.

We now claim that $W(k) = H(k, r)$. This is clear since

$h_{ij} = 1$ iff R_i^* is contained in R_j,
iff $\bar{R}_i^*$ is contained in $\bar{R}_j$ (since R_s is a subset of $\{-k + 2, -k + 3, \cdots, 0\}$ for every s),
iff $\bar{s} \in \bar{R}_i - \{\bar{0}\}$ implies $\overline{s+1} \in \bar{R}_j$ (by definition of R_i^*),
iff $\bar{s} \in \bar{R}_i - \{\bar{0}\}$ implies $\bar{s} \notin \bar{C}_j$ (by definition of C_j, and working modulo k),
iff $\bar{R}_i$ and $\bar{C}_j$ are disjoint (since $\bar{0} \notin \bar{C}_s$ for any s),
iff $w(k)_{ij} = 1$

and hence $W(k) = H(k, r)$.

Second, consider $W(n)$, $n \geqq k$. By Theorem 5.5 (ii)

$$\begin{aligned} w(n+1)_{ij} &= v_{n+1,k}(X_{ij}) \\ &= \sum_{*} v_{n,k}(X_{sj}) \quad \text{(where } \textstyle\sum_{*} \text{ represents the sum over all } s \in \{1, 2, \cdots, t\} \text{ such} \\ &\qquad\qquad \text{that } R_i^* \text{ is a subset of } R_s\text{),} \\ &= \sum_{s=1}^{t} h_{is} v(n)_{sj}, \end{aligned}$$

i.e., $W(n+1) = H(k, r).W(n) = H(k, r)^{n-k+1}$ for every $n \geqq k$.

Now suppose $n = 2k - 2$. Then, by Theorem 5.6, and because of the chosen labeling, $u_k(X_{ij}) = v_{2k-2,k}(X_{ji}) = w(2k-2)_{ji}$. Thus, by Theorem 5.4 we have

$$\begin{aligned} a(n,k,r) &= \sum_{E \in \mathbf{E}(k,r)} u_k(E) v_{n,k}(E) \\ &= \sum_{i=1}^{t} \sum_{j=1}^{t} u_k(X_{ij}).v_{n,k}(X_{ij}) \\ &= \sum_{i=1}^{t} \sum_{j=1}^{t} w(2k-2)_{ji}.w(n)_{ij} \\ &= \mathrm{Trace}\,(W(n).W(2k-2)) \\ &= \mathrm{Trace}\,(H(k,r)^{n-k+1}.H(k,r)^{k-1}) \\ &= \mathrm{Trace}\,(H(k,r)^n). \end{aligned} \qquad \square$$

Corollaries 4.5 and 4.6(i) are immediate from Theorem 4.4. We now prove the asymptote for $a(n, k, r)$ given in Corollary 4.6(ii).

Proof of Corollary 4.6(ii).

$$\begin{aligned} a(n,k,r)/(q_s)^n &= \sum_{i=1}^{t} \left(\frac{q_i}{q_s}\right)^n \\ &= \sum_{\substack{i=1 \\ i \neq s}}^{t} \left(\frac{q_i}{q_s}\right)^n + 1. \end{aligned}$$

Now

$$\left| \sum_{\substack{i=1 \\ i \neq s}}^{t} \left(\frac{q_i}{q_s}\right)^n \right| \leqq \sum_{\substack{i=1 \\ i \neq s}}^{t} \left|\frac{q_i}{q_s}\right|^n \leqq (t-1).d^n$$

where $d = \max_{i \neq s}(|q_i/q_s|) < 1$. Finally, note that $(t-1).d^n$ can be made arbitrarily small given sufficiently large n, and the result follows. □

To establish 4.7 and 4.8 we need to examine the matrix $H(k, 1)$. In fact we have

LEMMA 5.7. *Suppose $r = 1$ and $k \geqq 2$. Then if the labeling $(R_1, R_2, \cdots, R_t)$ is chosen so that $R_i = \{1 - i\}$, then $H(k, 1)$ is the $k-1$ by $k-1$ matrix*

$$\begin{bmatrix} 1 & 1 & 1 & \cdots\cdots & 1 \\ & & & & 0 \\ & & & & 0 \\ & & I_{k-2} & & \vdots \\ & & & & 0 \end{bmatrix}$$

where I_{k-2} is the $k-2$ by $k-2$ identity matrix.

Proof. First note that $\{0\}^*$ is empty, and hence $h_{1j} = 1$ for every j. Second, note that if $i < 0$, then $\{i\}^* = \{i + 1\}$, and so if $i > 1$ then $h_{ij} = 1$ if and only if $j = i - 1$. □

Hence, for the case $r = 1$, $H(k, 1)$ is already in Frobenius normal form, and as in [9, p. 297] the characteristic equation of $H(k, 1)$ is

$$x^{k-1} - \sum_{i=0}^{k-2} x^i = 0.$$

This gives Result 4.7 as an immediate corollary. We can also now prove the final result from § 4.

Proof of Theorem 4.8. As before let $r = 1$ and $k \geqq 2$. Then we claim that if $1 \leqq i \leqq k - 1$, then $H(k, r)^i =$

$$\begin{bmatrix} \multicolumn{2}{c}{\mathbf{d}_i} \\ \multicolumn{2}{c}{\mathbf{d}_{i-1}} \\ \multicolumn{2}{c}{\vdots} \\ \multicolumn{2}{c}{\mathbf{d}_1} \\ I_{k-i-1} & O_{k-i-1,i} \end{bmatrix}$$

where I_{k-i-1} is the $(k - i - 1)$ by $(k - i - 1)$ identity matrix, $O_{k-i-1,i}$ is the $(k - i - 1)$ by i all-zero matrix and $\mathbf{d}_i = (d_{i1}, d_{i2}, \cdots, d_{i(k-1)})$ satisfies $d_{ij} = 2^{i-1}$, $1 \leqq j \leqq k - i$.

By Lemma 5.7 this is clearly true for $i = 1$, and by induction (and by examination of $H(k, 1)$) we need only observe that

$$d_{ij} = \sum_{s=1}^{i-1} d_{sj} + \begin{cases} 1 & \text{if } j \leqq k - i, \\ 0 & \text{if } j > k - i. \end{cases}$$

Hence, if $j \leqq k - i$,

$$\begin{aligned} d_{ij} &= \sum_{s=1}^{i-1} 2^{s-1} + 1 \quad \text{(by the inductive hypothesis)} \\ &= 2^{i-1}. \end{aligned}$$

Thus,

$$\begin{aligned} \text{Trace } (H(k, r)^i) &= d_{i1} + d_{(i-1)2} + \cdots + d_{1i} \quad (i \leqq k - 1) \\ &= 2^{i-1} + 2^{i-2} + \cdots + 2^0 \\ &= 2^i - 1. \end{aligned}$$

□

Note also that $a(1, k, r) = 1$ for every k and r since R^* is contained in R iff $R = \{-r + 1, -r + 2, \cdots, 0\}$, and thus $H(k, r)$ always has a unique nonzero diagonal entry.

6. Tabulations of computed values. The papers of Metropolis, Stein and Stein [9], and Minc [10], contain extensive tables of values for $a(n, k)$ for $k \leqq 9$; [9] also contains tables of the characteristic equations for $H(k, r)$ and approximate values for the maximal eigenvalue of $H(k, r)$, again for $k \leqq 9$.

Using Theorem 4.4, together with a set of multiprecision routines written by Dave Levin running on a VAX = 11/750 minicomputer, we have been able to verify all the existing tabulations of $a(n, k)$ and $a(n, k, r)$, and to also produce the following tables of values for $k = 10$, 11 and 12 and $1 \leqq n \leqq 50$. (See Tables 1-3.) Note that, as in the remarks following Corollary 4.6 in § 4, we define $a(n, k, r)$ to be the trace of $H(k, r)^n$ for every $n \geqq 1$, and, in the natural way, we define $a(n, k)$ to be the sum of the $a(n, k, r)$ for every $n \geqq 1$.

TABLE 1
$a(n, 10)$ $(1 \leqq n \leqq 50)$

n	$a(n, 10)$
1	10
2	50
3	226
4	962
5	3840
6	16130
7	65698
8	258690
9	986410
10	3628800
11	14684570
12	59216642
13	238282730
14	957874226
15	3850864416
16	15498424578
17	62494094138
18	252579461906
19	1023207993178
20	4152609019392
21	16866126115498
22	68562634725426
23	278965798055154
24	1136049057102978
25	4630217243007040
26	18885572768497186
27	77080942110390418
28	314787782093356610
29	1286217554205276682
30	5257934625513024000
31	21503218756525334970
32	87975626996492343810
33	360060541514858306810
34	1474102716437359422226
35	6036778093871268296928
36	24728373540667369577474
37	101318258384798761261866
38	415213810742569786850322
39	1701918744817772671844282
40	6977191966118035882693120
41	28608161263286199980584138
42	117316730697716871569616818
43	481154617504945351421631490
44	1973597676853638993657364034
45	8096120287083522358723474560
46	33215073534422084882289815106
47	136279156753579083576867246210
48	559185646824298651823816588034
49	2294624949149162154512316665962
50	9416588798300969653474145747200

7. Developments of the basic problem. The determination of $a(n, k)$ is only one of many problems associated with the design of a sliding window time element scrambler of the type described in § 3 above. There is also the fundamental problem of choosing n and k, and designing the method to be used to select permutations from $A(n, k)$.

TABLE 2
$a(n, 11)\,(1 \leqq n \leqq 50)$

n	$a(n, 11)$
1	11
2	61
3	299
4	1393
5	6331
6	27949
7	126095
8	554177
9	2368847
10	9864101
11	39916800
12	176214841
13	775596313
14	3407118041
15	14951584189
16	65598500129
17	287972983669
18	1265785879297
19	5573449326001
20	24588660672953
21	108681408827381
22	481065936784384
23	2130831306657527
24	9445455128274737
25	41902710214254531
26	186040589545320129
27	826626380784149855
28	3675606432528120601
29	16354817596119737239
30	72817892293114361249
31	324404970589895718419
32	1446036425685642910913
33	6449154750576695662848
34	28777322874980997201469
35	128473548843752900117725
36	573831697082734230011665
37	2564217910410345862799157
38	11463508074975657944297053
39	51270268001103972812908657
40	229399692125416838094166177
41	1026818034189449323389052049
42	4597927569350275420770702533
43	20596506835524484240745827169
44	92295992963140763623590913024
45	413737754483439976252567341907
46	1855307333069535348229092448661
47	8322436742793852726661366713051
48	37344337184202486272125701583553
49	167623315461313026160891570970211
50	752619449962479689980066343390501

As before we let

$$A(n,k)=\{\pi\in S_n : \overline{i\pi}\in\{\bar{i},\overline{i+1},\cdots,\overline{i+k-1}\}\text{ for every } i\}.$$

Another secondary problem, similar to the $a(n, k)$ evaluation problem, concerns choosing

TABLE 3
$a\,(n, 12)\ (1 \leqq n \leqq 50)$

n	$a\,(n, 12)$
1	12
2	72
3	384
4	1944
5	9812
6	46080
7	227680
8	1100680
9	5199648
10	24011832
11	108505112
12	479001600
13	2290792932
14	10927434464
15	52034548064
16	247524019720
17	1177003136892
18	5598118158336
19	26647751359904
20	127007092256024
21	606269105086336
22	2898753047375312
23	13880706183899752
24	66544727442343936
25	319198916117248012
26	1532071808279181592
27	7358305929283036608
28	35363678926464144632
29	170062683110076661012
30	818309438846696002560
31	3939711747851871915248
32	18977103341489089532424
33	91452381430150298900000
34	440902914787573840187976
35	2126473158349980849520200
36	10259701680625467679872000
37	49517433552724675102157540
38	239067514640241762853861328
39	1154549828245379314130268192
40	5577319090541480294809775880
41	26949490191171589347220311676
42	130250684430090783496906489856
43	629660737886339608173390416560
44	3044553776812595993002687353336
45	14723969563417452202403843439488
46	71220434757273136282267411587712
47	344554065382463547747151575797784
48	1667163251724747083829231695497216
49	8067930334499348958454566728595916
50	39048557417232324389011734475683432

permutations suitable for use from $A(n, k)$. Clearly not every permutation in $A(n, k)$ is suitable for use as a scrambling pattern; consider the permutation $\pi \in S_n$ which satisfies $i\pi = i - 1$ $(2 \leqq i \leqq n)$ and $1\pi = n$. Then $\pi \in A(n, k)$ for every $k \geqq 1$, but the transmitted speech enciphered using π will, in effect, not be permuted at all, and a device using such a permutation will offer no security at all.

The basic problem is what is commonly known as *residual intelligibility*. This term refers to the amount of intelligible information remaining in the analogue signal after it has been scrambled. Clearly, different permutations from $A(n, k)$ will have different residual intelligibilities, and it is thus desirable to have some method of choosing permutations from $A(n, k)$ which leave the minimum residual intelligibility.

In order to assess the level of residual intelligibility associated with a permutation, it is necessary to perform a large number of experiments to try to assess the amount of decipherable information remaining in speech after encryption using the permutation. Such experiments have been performed, and the results of these experiments have led us to conclude that the most important extra criterion that a permutation $\pi \in A(n, k)$ should satisfy in order to minimise the residual intelligibility is that

$$\overline{i\pi+1} \neq \overline{(i+1)\pi}, \quad 1 \leqq i \leqq n-1 \quad \text{and} \quad \overline{n\pi+1} \neq \overline{1\pi}.$$

This ensures that no two originally consecutive segments remain consecutive after encryption.

Thus, if we let

$$B(n,k) = \{\pi \in A(n,k) : \overline{i\pi+1} \neq \overline{(i+1)\pi}, 1 \leqq i \leqq n-1 \quad \text{and} \quad \overline{n\pi+1} \neq \overline{1\pi}\}$$

and $b(n, k) = |B(n, k)|$, then choosing permutations from $B(n, k)$ considerably reduces the probability of π leaving a high level of residual intelligibility in the scrambled speech. For a more detailed description of the experimental results and permutation evaluation procedures (see [3] and [4]).

Once we have made this definition, it is clearly important that some estimate be obtained for the size of $b(n, k)$. However, few results appear to exist on this problem, and the following summarises the results currently known to the authors.

THEOREM 7.1. (i) $b(n, 1) = b(n, 2) = 0$ *for every* n,

(ii) $b(n, 3) = b(n-2, 3) + b(n-3, 3)$, $n \geqq 6$, $b(3, 3) = 3$, $b(4, 3) = 2$, $b(5, 3) = 5$,

(iii) $b(n, 4) = 2b(n, 3)$, $n \geqq 4$,

(iv) $b(n, n) = n.\sum_{i=1}^{n-1} (-1)^{i-1}.a(n-i, n-i-1)$, $n \geqq 2$.

Theorem 7.1(i) is trivial. Parts (ii) and (iii) have been obtained independently by Dr. Keith Lloyd and the authors. Part (iv) is based on a recurrence relation due to Stacey [8], which says that $b(n+3, n+3) = n.b(n+2, n+2) + 2.(n+1).b(n+1, n+1) + (n+1).b(n, n)$. The solution to this recurrence to give (iv) can be found in [7, Ex. 15.5.10]. For further references to (iv) see also [16, Exercise 21, p. 160] and [16, Exercise 8, p. 172]. We now give a proof of (ii) and (iii).

In order to prove these two results, we first need some preliminary definitions. Let

$$B(n,k,r) = \{\pi \in B(n,k) : \pi \in A(n,k,r)\}.$$

As for Lemmas 4.2 and 4.3 we immediately have

LEMMA 7.2.

$$b(n,k) = \sum_{r=0}^{k-1} b(n,k,r).$$

Proof. Immediate from the definition. □

LEMMA 7.3. (i) $b(n, k, r) = b(n, k, k-1-r)$, $0 \leqq r \leqq k-1 \leqq n-1$.

(ii) $b(n, k, 0) = b(n, k, k-1) = 0$, $0 \leqq k-1 \leqq n-1$.

Proof. (i) As for the proof of Lemma 4.3(i), we define the function ϕ_k which maps S_n into S_n by

$$\phi_k(\pi) \text{ maps } i \text{ to } (n+1) - s\pi \text{ where } s \in \{1, 2, \cdots, n\} \text{ and } \bar{s} = \overline{-i-k+2}.$$

We claim that ϕ_k is a one-to-one mapping from $B(n, k, r)$ into $B(n, k, k - 1 - r)$. This will establish the result. By the proof of Lemma 4.3(i) we have shown that ϕ_k is one to one and that if $\pi \in B(n, k, r)$ then $\phi_k(\pi) \in A(n, k, k - 1 - r)$; hence we need only show that $\phi_k(\pi) \in B(n, k)$ in order to establish the above claim, and thence the desired result. Now if i and j satisfy $\overline{i+1} = \bar{j}$, $i, j \in \{1, 2, \cdots, n\}$, then $i\phi_k(\pi) + 1 = n + 2 - t\pi$, and $j\phi_k(\pi) = n + 1 - s\pi$, where, by definition, $\overline{s+1} = \bar{t}$. But $\pi \in B(n, k, r)$, and because s and t satisfy $\overline{s+1} = \bar{t}$, we know $\overline{s\pi+1} \neq \bar{t}$. Hence $\overline{i\phi_k(\pi)+1} \neq \overline{j\phi_k(\pi)}$, and thus $\phi_k(\pi) \in B(n, k)$.

(ii) This part is trivial. □

We can now give the following lemma.

LEMMA 7.4. $B(n, k, k - 2) = B(n, 3, 1)$, $k \geqq 3$.

Proof. We first show that if $1 \leqq r < k < n$, then $A(n, k - 1, r - 1)$ is a subset of $A(n, k, r)$. Suppose that $\pi \in A(n, k - 1, r - 1)$. Then, by definition, $\pi \in A(n, k - 1)$ and hence $\pi \in A(n, k)$. Thus, by Lemma 4.1, we need only show that $|X_k(\pi, i)| = r$ for some $i \in \{1, 2, \cdots, n\}$.

Now $\pi \in A(n, k - 1, r - 1)$, and hence $|X_{k-1}(\pi, n - k + 3)| = r - 1$. By definition, $X_{k-1}(\pi, n - k + 3)$ is a subset of $X_k(\pi, n - k + 2)$, and

$$X_k(\pi, n-k+2) - X_{k-1}(\pi, n-k+3) = \{(n-k+2)\pi\},$$

since $\pi \in A(n, k - 1)$. Thus:

$$|X_k(\pi, n-k+2)| = r - 1 + |\{(n-k+2)\pi\}| = r,$$

and hence $A(n, k - 1, r - 1)$ is a subset of $A(n, k, r)$. This immediately implies that $B(n, 3, 1)$ is a subset of $B(n, k, k - 2)$, $k \geqq 3$.

We now show that $B(n, k, k - 2)$ is a subset of $B(n, 3, 1)$, $k \geqq 3$, and the result follows.

Clearly, if $k = 3$, then the claim is automatically true, and so we suppose $k \geqq 4$. Now choose $\pi \in B(n, k, k - 2)$, and suppose $\pi \notin B(n, 3)$, i.e., suppose there exists an $h \in \{1, 2, \cdots, n\}$ for which $\overline{h\pi} = \overline{h+s}$, where $3 \leqq s \leqq k - 1$.

Now, by definition, $|X_k(\pi, i)| = k - 2$, for every $i \in \{1, 2, \cdots, n\}$. Let $x, y \in \{1, 2, \cdots, n\}$ satisfy $\bar{x} = \overline{h-k+3}$ and $\bar{y} = \overline{h-k+4}$. Since

$$\overline{h\pi} \notin \{\bar{h}, \overline{h+1}, \overline{h+2}\}$$

we have: $h\pi \notin X_k(\pi, x)$ and $h\pi \notin X_k(\pi, y)$. But

$$|X_k(\pi, x)| = |X_k(\pi, y)| = k - 2,$$

and hence if $u, v \in \{1, 2, \cdots, n\}$ satisfy $\bar{u} = \overline{h+1}$ and $\bar{v} = \overline{h+2}$ then $u\pi = u$ and $v\pi = v$. But since $\bar{v} = \overline{u+1}$ this contradicts the definition of $B(n, k)$ and hence $\pi \in B(n, 3)$. The result now follows by our observing that $B(n, 3) = B(n, 3, 1)$, since $B(n, 3, 0)$ and $B(n, 3, 2)$ are empty by Lemma 7.3(ii). □

Now since $b(n, 4) = b(n, 4, 1) + b(n, 4, 2)$ (by Lemmas 7.2 and 7.3(ii)), and since $b(n, 4, 1) = b(n, 4, 2)$ (by Lemma 7.3(i)), we know that $b(n, 4) = 2b(n, 4, 2)$. But $b(n, 4, 2) = b(n, 3, 1)$ (by Lemma 7.4), and hence $b(n, 4) = 2b(n, 3, 1)$, establishing Theorem 7.1(iii). It remains for us to prove the recurrence of Theorem 7.1(ii), noting that the initial values of $b(n, 3)$ for $n \leqq 5$ can be verified by hand.

Proof of Theorem 7.1(ii). We first introduce some notation.

Suppose $n \geqq 3$. Let

$$Q(n) = \{\pi \in B(n, 3) : 1\pi = 1\} \quad \text{and} \quad q(n) = |Q(n)|.$$

Also define

$$Q_1(n) = \{\pi \in Q(n) : 3\pi = 3\} \quad \text{and} \quad q_1(n) = |Q_1(n)|,$$

$$Q_2(n) = \{\pi \in Q(n) : 3\pi \neq 3\} \quad \text{and} \quad q_2(n) = |Q_2(n)|.$$

Then $Q(n)$ is equal to the disjoint union of $Q_1(n)$ and $Q_2(n)$, and we have

(1) $$q(n) = q_1(n) + q_2(n), \qquad n \geqq 3.$$

We also need the notion of a displacement vector. Choose $\pi \in S_n$, and let $\mathbf{d} = (d_1, d_2, \cdots, d_n)$ satisfy:

$$d_i \in \{0, 1, \cdots, n-1\} \quad \text{and} \quad \bar{d}_i = \overline{i\pi - i} \quad \text{for every } i \in \{1, 2, \cdots, n\}.$$

Then we call $\mathbf{d}$ the displacement vector of π. Note that permutations in $Q_1(n)$ and $Q_2(n)$ have displacement vectors of the form $(0, 2, \cdots)$ and $(0, 1, 2, \cdots)$, respectively.

Now suppose $n \geqq 5$. We define the mapping ϕ_1 from $Q_1(n)$ into $Q(n-2)$ as follows. If $\pi \in Q_1(n)$ has displacement vector

$$(0, 2, d_3, d_4, \cdots, d_n)$$

then let $\phi_1(\pi)$ be the permutation having displacement vector

$$(d_3, d_4, \cdots, d_n).$$

It is straightforward to show that ϕ_1 is well defined and both one to one and onto. We have thus shown:

(2) $$q_1(n) = q(n-2), \qquad n \geqq 5.$$

Next suppose $n \geqq 6$. We define the mapping ϕ_2 from $Q_2(n)$ into $Q(n-3)$ as follows. If $\pi \in Q_2(n)$ has displacement vector

$$(0, 1, 2, d_4, d_5, \cdots, d_n)$$

then let $\phi_2(\pi)$ be the permutation having displacement vector

$$(d_4, d_5, \cdots, d_n).$$

It is straightforward to show that ϕ_2 is well-defined and both one to one and onto. We have thus shown:

(3) $$q_2(n) = q(n-3), \qquad n \geqq 6.$$

Next suppose $n \geqq 4$ and define a third mapping ϕ_{12} from $Q_1(n-1)$ into $Q_2(n)$ as follows. If $\pi \in Q_1(n-1)$ has displacement vector

$$(0, 2, d_3, d_4, \cdots, d_{n-1})$$

then let $\phi_{12}(\pi)$ be the permutation having displacement vector

$$(0, 1, 2, d_3, d_4, \cdots, d_{n-1}).$$

Again it is straightforward to show that ϕ_{12} is well defined and both one to one and onto. We then have

(4) $$q_1(n-1) = q_2(n), \qquad n \geqq 4.$$

Finally suppose $n \geqq 3$. If $\mathbf{d}$ is the displacement vector of $\pi \in B(n, 3)$, and if $\pi^* \in B(n, 3)$ has displacement vector $\mathbf{d}^* = (d_{s+1}, d_{s+2}, \cdots, d_n, d_1, d_2, \cdots, d_s)$, then we call π^* the s-fold cyclic shift of π. We now let

$$Q_{11}(n) = \{\pi^* : \pi^* = \text{the 1-fold cyclic shift of some } \pi \in Q_1(n)\},$$

$$Q_{21}(n) = \{\pi^* : \pi^* = \text{the 1-fold cyclic shift of some } \pi \in Q_2(n)\},$$

$$Q_{22}(n) = \{\pi^* : \pi^* = \text{the 2-fold cyclic shift of some } \pi \in Q_2(n)\}.$$

It is straightforward to show that all elements of $Q_{11}(n)$, $Q_{21}(n)$ and $Q_{22}(n)$ have displacement vectors of the forms: $(2, 0, \cdots, 0)$, $(1, 2, 0, \cdots)$ and $(2, 0, \cdots, 1)$, respectively. Hence the five sets

$$Q_1(n), Q_{12}(n), Q_2(n), Q_{21}(n), Q_{22}(n)$$

are all disjoint; moreover, every element of $B(n, 3)$ is in one of these sets. This immediately gives

$$b(n, 3) = 2q_1(n) + 3q_2(n), \qquad n \geqq 3. \tag{5}$$

We can now combine the above results to obtain the desired recurrence. Suppose $n \geqq 6$. Then:

$$\begin{aligned} b(n,3) &= 2q_1(n) + 3q_2(n) \quad \text{by (5)} \\ &= 2q(n-2) + 3q(n-3) \quad \text{by (2) and (3)} \\ &= 2q_1(n-2) + 2q_2(n-2) + 3q_1(n-3) + 3q_2(n-3) \quad \text{by (1)} \\ &= 2q_1(n-2) + 3q_2(n-2) + 2q_1(n-3) + 3q_2(n-3) \quad \text{by (4)} \\ &= b(n-2,3) + b(n-3,3) \quad \text{by (5).} \end{aligned}$$ □

Acknowledgments. The authors would like to acknowledge the assistance and encouragement of Racal Comsec Ltd. and Racal Research Ltd. The authors would also like to thank Dr. Keith Lloyd for his help in discovering the history of the problem.

REFERENCES

[1] H. J. BEKER, *Analogue speech security systems*, in Cryptography: Proc. Workshop on Cryptography, Burg Feuerstein 1982, Lecture Notes in Comput. Sci., 149, Springer-Verlag, Berlin, New York, 1983, pp. 130–146.

[2] ———, *Options available for speech encryption*, Radio Electron. Engineer, 54 (1984), pp. 35–40.

[3] H. J. BEKER AND F. C. PIPER, *Secure Speech Communications*, Academic Press, London, 1985.

[4] A. J. BROMFIELD AND C. J. MITCHELL, *Permutation selector for a sliding window time element scrambler*, preprint.

[5] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[6] I. KAPLANSKY AND J. RIORDAN, *The Problème des Ménages*, Scripta Math., 12 (1946), pp. 113–124.

[7] W. LEDERMANN, (ed.), *Handbook of Applicable Mathematics, Volume* 5: *Geometry and Combinatorics*, John Wiley, New York, 1985.

[8] E. K. LLOYD, Private communication, May 1982.

[9] N. METROPOLIS, M. L. STEIN AND P. R. STEIN, *Permanents of cyclic* (0, 1) *matrices*, J. Combin. Theory, 7 (1969), pp. 291–321.

[10] H. MINC, *Permanents of* (0, 1)-*Circulants*, Canad. Math. Bull., 7 (1964), pp. 253–263.

[11] ———, *Permanents*, Addison-Wesley, Reading, MA, 1978.

[12] C. J. MITCHELL AND F. C. PIPER, *A classification of time element speech scramblers*, J. Institut. Electron. Radio Engineers, 55 (1985), pp. 391–396.

[13] W. O. J. MOSER, *The number of very reduced* 4 × *n Latin rectangles*, Canad. J. Math., 19 (1967), pp. 1011–1017.

[14] J. RIORDAN, *Discordant permutations*, Scripta Math., 20 (1954), pp. 14–23.

[15] J. TOUCHARD, *Permutations discordant with two given permutations*, Scripta Math., 19 (1953), pp. 109–119.

[16] A. TUCKER, *Applied Combinatorics*, John Wiley, New York, 1980.

[17] L. G. VALIANT, *The complexity of computing the permanent*, Theoret. Comput. Sci., 8 (1979), pp. 189–201.

[18] ———, *The complexity of enumeration and reliability problems*, SIAM J. Comput., 8 (1979), pp. 410–421.

[19] E. G. WHITEHEAD, *Four-discordant permutations*, J. Austral. Math. Soc. Ser. A, 28 (1979), pp. 369–377.

[20] K. YAMAMOTO, *Structure polynomial of Latin rectangles and its application to a combinatorial problem*, Mem. Fac. Sci. Kyusyu Univ. Ser. A, 10 (1956), pp. 1–13.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

005

MARKOV MAPS AND THE SPECTRAL RADIUS OF 0-1 MATRICES*

W. BYERS† AND A. BOYARSKY‡

Abstract. Let $\mathscr{A}$ denote the set of $n \times n$ 0-1 matrices, $n = 1, 2, \cdots$, where the nonzero entries in each row are contiguous. Let $A, B \in \mathscr{A}$ be irreducible and have the same shape. The main result states that under certain conditions A and B must have the same spectral radius.

Key words. nonnegative matrices, Markov maps, spectral radius

AMS(MOS) subject classifications. Primary 15A15; secondary 26-00

1. Introduction. The methods and results of linear algebra have played an important role in ergodic theory [9], [10]. In the theory of Markov chains, for example, the Frobenius–Perron Theorem is used to determine the longterm behavior of random processes. In [4] ideas from the ergodic theory of transformations from [0, 1] into [0, 1] were used to derive conditions which guarantee the irreducibility and primitivity of large matrices from the irreducibility and primitivity of smaller ones. The key tool is the application of a class of piecewise linear Markov maps, whose dynamical behavior can be fully understood by using the tools of linear algebra [1], [2], [4], [6], [7], [8]. These maps serve as links between ergodic theory and linear algebra and allow results to pass both ways. The purpose of this paper is to use the ergodic theory of transformations to present conditions under which two different sized 0-1 matrices have the same spectral radius. In particular, we shall study a class of 0-1 matrices which are induced by piecewise linear Markov maps. A simple example of such a matrix is the following:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

where we think of A as being induced by the piecewise linear Markov map f: $[0, 1] \to [0, 1]$, shown in Fig. 1. The intervals $I_1 = [0, 1/2)$ and $I_2 = [1/2, 1]$ form a partition of $[0, 1]$. If we define $A = (a_{ij})$ by $a_{ij} = 1$ if $f(I_i) \supseteq I_j$ and 0 otherwise, we obtain the matrix A above. We shall say that the map f induces the matrix A with respect to the partition $Q = \{0, 1/2, 1\}$.

Now let f be as above, but consider a different partition of [0, 1], namely, $Q_1 = \{0, 1/8, 1/4, 1/2, 1\}$. Let $J_1 = [0, 1/8)$, $J_2 = [1/8, 1/4)$, $J_3 = [1/4, 1/2)$ and $J_4 = [1/2, 1]$, as shown in Fig. 2. Then, using the above definition of (a_{ij}), f induces the following matrix:

$$A_1 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

with respect to the partition Q_1. In a certain sense both A and A_1 inherit the shape of the map f, and they both have spectral radius 2. The main result of this paper (§ 2)

* Received by the editors February 6, 1985; accepted for publication (in revised form) September 9, 1986.

† Department of Mathematics, Loyola Campus, Concordia University, Montreal, Canada H4B 1R6.
The work of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A-5325.

‡ Department of Mathematics, Loyola Campus, Concordia University, Montreal, Canada H4B 1R6.
The work of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A-9072.

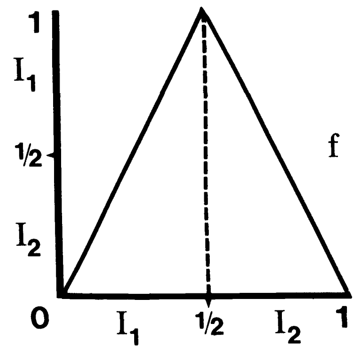


FIG. 1

characterizes the matrices induced by a general piecewise linear Markov map (not necessarily continuous) which have the same spectral radius. This result is useful in finding the spectral radius of large matrices. Given a large matrix, B, of the type we define below, we present conditions under which there exist smaller matrices, A, in the same class which have the same spectral radius. In § 3 we show that if the characteristic polynomial of A, $p_A(x)$, is irreducible over the integers, then it is a factor of $p_B(x)$. In § 4 we consider two $n \times n$ 0-1 matrices A and B which are induced by two piecewise linear Markov maps τ and γ on the same partition of $[0, 1]$. We derive a relation between the spectral radius of $\lambda A + (1 - \lambda)B$, $0 < \lambda < 1$, in terms of the spectral radius of A and the spectral radius of B.

2. Notation and preliminary results.

DEFINITION 1. Let $I = [a, b]$ be a closed interval and let

$$Q = \{a = a_0 < a_1 < \cdots < a_n = b\}$$

be a set of partition points of I. We say that $f: I \to I$ is a piecewise-continuous Markov map with respect to the partition points Q if (1) f is strictly monotonic and continuous on each subinterval $I_i = (a_{i-1}, a_i)$ $i = 1, \cdots, n$ and (2) both the right and left limits $f(a_i^+) = \lim_{x \downarrow a_i} f(x)$ and $f(a_i^-) = \lim_{x \uparrow a_i} f(x)$ are elements of Q. Let $\mathscr{C}$ denote the class of Markov maps which are piecewise linear with respect to their defining partition, i.e., f is linear on I_i for $i = 1, \cdots, n$.

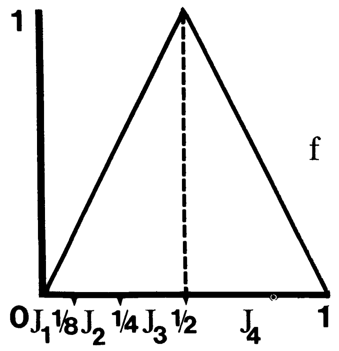


FIG. 2

Let $f \in \mathscr{C}$ and let $\{I_i\}_{i=1}^{n}$ be the intervals of the partition with respect to which f is Markov. Then f induces an $n \times n$ matrix $M = M_\tau$ defined as follows:

$$m_{ij} = \frac{1}{|f'_i|} d_{ij}$$

where $f'_i = df/dx|_{I_i}$ and $d_{ij} = 1$ if $f(I_i) \supset I_j$ and 0 otherwise. Thus, all nonzero entries of each row in M are contiguous (no zero entries between nonzero entries) and equal with common value $1/|f'_i|$.

LEMMA 1 [1]. *Let $f \in \mathscr{C}$. Then $M = M_f$ has* 1 *as its spectral radius. If M is also irreducible, then the algebraic and geometric multiplicity of the eigenvalue* 1 *are also* 1.

Note that the row sums of M are not necessarily 1. If they were the result would follow immediately from the Perron–Frobenius Theorem.

DEFINITION 2. Let $f \in \mathscr{C}$. Then f induces the 0-1 matrix $A = A_f$, where (a_{ij}) is given by $a_{ij} = d_{ij}$. Let $\mathscr{B}$ denote the class of 0-1 matrices induced by all $f \in \mathscr{C}$.

Clearly, given $f \in \mathscr{C}$, A_f can be obtained from M_f by replacing every nonzero entry in M_f by 1. The following result is proved in [2]:

LEMMA 2. *Let $A \in \mathscr{B}$ be irreducible and let it have spectral radius $\lambda > 1$. Then there exists $f \in \mathscr{C}$, i.e., a piecewise linear Markov map f*: [0, 1] → [0, 1], *having constant slope* λ, *such that $M_f = A/\lambda$.*

Let $\rho(A)$ denote the spectral radius of A. Then it follows from Lemma 1 that $\rho(M_f) = \rho(A/\lambda) = 1$, i.e., $\rho(A) = \lambda$.

3. Matrices shaped like maps. We wish to characterize the set of matrices $A \in \mathscr{B}$ which can be generated from a single map $f \in \mathscr{C}$ by means of different partitions Q.

Take, for example, the map $f \in \mathscr{C}$ defined by

$$f(x) = \begin{cases} x + \frac{1}{2}, & 0 \leqq x < \frac{1}{2}, \\ 2x - 1, & \frac{1}{2} \leqq x \leqq 1, \end{cases}$$

as shown in Fig. 3. With respect to $Q_1 = \{0, \frac{1}{2}, 1\}$, f induces the matrix $A_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ whereas with respect to $Q_2 = \{0, 1/4, 1/2, 5/8, 3/4, 1\}$ f induces the following matrix:

$$A_2 = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

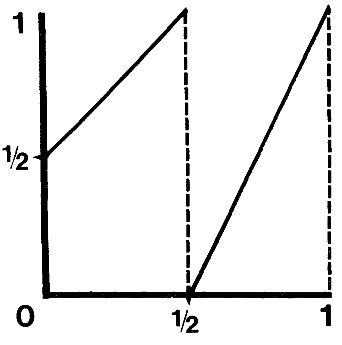


FIG. 3

The crucial factor in defining the matrices A_i is the behavior of f on the invariant sets Q_1 and Q_2. If we let $a_0 = 0$, $a_1 = 1/2$ and $a_2 = 1$ we may define a finite map σ by the following table:

i	0	1^-	1^+	2
$\sigma(i)$	1	2	0	2

where $f(a_i^-) = a_{\sigma(i^-)}$ and $f(a_i^+) = a_{\sigma(i^+)}$ which gives enough information to define the matrix A_1. Similarly the action of τ on Q_2 is captured by the following table.

i	0	1	2^-	2^+	3	4	5
$\sigma'(i)$	2	4	5	0	1	2	5

.

Notice that σ' has an invariant set $\{0, 2^-, 2^+, 5\}$ on which the action of σ' is isomorphic to that of σ. We shall show that this forces the spectral radius of A_2 to equal that of A_1 (in this case $(1 + \sqrt{5})/2$).

In general let $T = T_n = \{0, 1, \cdots, n\}$ and let (σ^-, σ^+) be a pair of maps from T into itself. If $\sigma^-(i) = \sigma^+(i)$, we denote their common value by $\sigma(i)$. We shall assume that neither σ^- nor σ^+ is constant on any adjacent pair $\{i - 1, i\}$. Of course any such pair (σ^-, σ^+) generates a piecewise linear Markov map $f \in \mathscr{C}$ from the interval $[0, n]$ into itself by setting $f(i^-) = \sigma^-(i)$ and $f(i^+) = \sigma^+(i)$ and defining f on $(i - 1, i)$ to be the unique linear map with $f((i - 1)^+) = \sigma^+(i - 1)$ and $f(i^-) = \sigma^-(i)$. The function f is continuous at i iff $\sigma^+(i) = \sigma^-(i) = \sigma(i)$. The pair (σ^-, σ^+) defines the matrix $A = A_\sigma \in \mathscr{B}$, which is the matrix defined in Definition 2 for the map f on the partition T. Conversely any map $f \in \mathscr{C}$ with partition $Q = \{x_0 < x_1 < \cdots < x_n\}$ defines the maps (σ^-, σ^+) from T_n into itself by $\sigma^-(i) = f(x_i^-)$ and $\sigma^+(i) = f(x_i^+)$.

Suppose that $S = \{0 = s_0 < s_1 < \cdots < s_m = n\}$ is some subset of T. *S is invariant under* (σ^-, σ^+) if $\sigma^-(S) \subset S$ and $\sigma^+(S) \subset S$. (σ^-, σ^+) is *piecewise monotonic on* S if

(i) S is invariant under (σ^-, σ^+),

(ii) $\sigma^-(i) = \sigma^+(i)$ $(=\sigma(i))$ for $i \in T - S$, and

(iii) $\sigma^+(s_{i-1}) < \sigma(s_{i-1} + 1) < \cdots < \sigma(s_i - 1) < \sigma^-(s_i)$ or
$\sigma^+(s_{i-1}) > \sigma(s_{i-1} + 1) > \cdots > \sigma(s_i - 1) > \sigma^-(s_i)$ for $i = 1, \cdots, m$.

If (σ^-, σ^+) is piecewise monotonic on $S \subset T$ we can define the pair (μ^-, μ^+) on T_n by setting $\mu^-(i) = j$ if and only if $\sigma^-(s_i) = s_j$ and $\mu^+(i) = j$ iff $\sigma^+(s_i) = s_j$. We could then generate the $m \times m$ matrix A_μ as above. Assume that A_μ is irreducible. It is not difficult to see that the spectral radius $\rho(A_\mu) = 1$ if and only if A_μ is a permutation matrix. It follows from the piecewise monotonicity of (σ^-, σ^+) on S that A_σ is also a permutation matrix and so $\rho(A_\sigma) = 1$. For this reason we shall assume that $\rho(A_\mu) > 1$ in the sequel. Then we can use Lemma 2 to find a map $f \in \mathscr{C}$ with slope $\pm\lambda$ where $\rho(A_\mu) = \lambda$. In fact in the proof of this lemma we construct a map of constant slope, f, which is Markov with respect to a partition of $[0, 1]$,

$$Q_S = \{0 = x_{s_0} < x_{s_1} < \cdots < x_{s_m} = 1\}$$

where f is strictly monotonic on the intervals $I_i = [x_{s_{i-1}}, x_{s_i}]$ $i = 1, \cdots, m$ and $f(x_{s_i}^-) = x_{\sigma^-(s_i)}$, $f(x_{s_i}^+) = x_{\sigma^+(s_i)}$. We shall refer to this f as being generated by (σ^-, σ^+). Our main result is the following:

THEOREM 1. *Suppose (σ^-, σ^+) is defined on the finite set T and is piecewise monotonic on $S \subset T$. If the induced matrices A_σ and A_μ are irreducible, then their spectral radii are equal.*

The proof follows from the following lemmas.

LEMMA 3. *Let (σ^-, σ^+) be piecewise monotonic on S. Then for each $t \in T - S$, the set of points in T but not in S, there exists a unique point $x_t \in [0, 1]$ with the same itinerary as t, i.e., if $\sigma^k(t) \in T - S$ for $k = 0, 1, \cdots, p - 1$, then $f^p(x_t) \in$ interior (I_i) if $s_{i-1} < \sigma^p(t) < s_i$ and $f^p(x_t) = x_{s_i}$, if $\sigma^p(t) = s_i$, where $f \in \mathscr{C}$ is generated by (σ^-, σ^+).*

Proof. Suppose that $s_{i-1} < t < s_i$. Then $\sigma(t)$ must be between $\sigma^+(s_{i-1})$ and $\sigma^-(s_i)$ because (σ^-, σ^+) is monotonic between s_{i-1} and s_i. If $s_{j-1} < \sigma(t) < s_j$, then s_{j-1} and s_j must be between $\sigma^+(s_{i-1})$ and $\sigma^-(s_i)$. Since f is a homeomorphism on I_i, we must have $f(I_i) \supset I_j$. On the other hand if $\sigma(t) = s_j$ for some j, we must have $x_{s_j} \in f(I_i)$ and so there is a unique point $x \in I_i$ with $f(x) = x_{s_j}$.

Now, again, if $t \in T - S$, either (a) $\sigma^k(t) \in T - S$ for all $k = 0, 1, 2, \cdots$ or else (b) there is a smallest integer p for which $\sigma^p(t) = s \in S$. In case (a) there exists for each $k = 0, 1, \cdots$ an integer n_k such that $s_{n_k - 1} < \sigma^k(t) < s_{n_k}$ and therefore by the above, a sequence of intervals $\{I_{n_k}\}_{k=0}^{\infty}$ with $f(I_{n_k}) \supset I_{n_{k+1}}$. By a standard result (cf. [5]) there exists a closed interval $J \subset I_{n_0}$ such that $f^k(J) \subset I_{n_k}$ for $k = 1, 2, \cdots$.

The fact that there is a unique point with the itinerary follows from the expansiveness of f (i.e., $|\text{slope}| = \lambda > 1$).

In case (b) we have $s_{n_{k-1}} < \sigma^k(t) < s_{n_k}$ for $k = 1, \cdots, p - 1$ and $\sigma^p(t) = s_j$. As above we have the interval $J \subset I_{n_0}$ with $f^k(J) \subset I_{n_k}$ $k = 0, 1, \cdots, p - 1$. In fact we can choose J so that $f^{p-1}(J) = I_{n_{p-1}}$. As above there is a point $x_0 \in I_{n_{p-1}}$ with $f(x_0) = x_{s_j}$. Since f is a homeomorphism on J there is a unique point $x_t \in J$ with $f^p(x_t) = x_{s_j}$.

LEMMA 4. *If the matrix A_σ is irreducible, then the correspondence $t \rightarrow x_t$ is injective.*

Proof. Suppose $x_{t_1} = x_{t_2}$ for $t_1 \neq t_2 \in T - S$. As in Lemma 2 we either have (a)

$$(*) \qquad s_{n_k - 1} < \sigma^k(t_i) < s_{n_k}, \qquad i = 1, 2,$$

holds for all $k = 0, 1, 2, \cdots$ or else (b) condition $(*)$ holds for $k = 0, 1, \cdots, p - 1$ and $\sigma^p(t_i) = s \in S$, $i = 1, 2$.

In case (b) the piecewise monotonicity of σ implies that $\sigma^{p-1}(t_1) = \sigma^{p-1}(t_2), \cdots, \sigma(t_1) = \sigma(t_2)$, $t_1 = t_2$. We complete the proof by showing that case (a) reduces to case (b) when A_σ is irreducible.

The irreducibility of A_σ implies that $a_{t_2,1}^k > 0$ for some k. Thus, there is a chain $a_{t_2,n_1} = 1 = a_{n_1,n_2} = \cdots = a_{n_{k-1},1}$. In terms of the function f this means that $f(I_{t_2}) \supset I_{n_1}$, $f(I_{n_1}) \supset I_{n_2}, \cdots, f(I_{n_{k-1}}) \supset I_1$. The piecewise monotonicity of σ between $\sigma^k(t_1)$ and $\sigma^k(t_2)$ (and of f between $f^k(x_{t_1})$ and $f^k(x_{t_2})$), given by $(*)$, means that this sequence of inclusions can be replaced by equalities. Thus $f^k(I_{t_2}) = I_1$ where $f^k|I_{t_2}$ is a homeomorphism. Thus either $f^k(x_{t_2}) = 0$ or else $t_1 = t_2 - 1$ and $f^k(x_{t_1}) = 0$. But $f^k(x_{t_i}) = 0$ implies that $\sigma^j(t_i) \in S$ for some $j \leqq k$ by Lemma 2. Thus case (a) reduces to (b).

LEMMA 5. *The correspondence $t \rightarrow x_t$ is order preserving.*

Proof. Suppose $t_1 < t_2$. The result is clear if one (or both) of t_1, t_2 lie in S or if x_{t_1}, x_{t_2} lie in the interiors of different intervals I_i. Lemma 4 implies that $f^k(x_{t_i})$, $i = 1, 2$, cannot lie in the interior of the same interval I_{n_k} for all $k = 0, 1, \cdots$. Thus the points $\sigma^k(t_i)$, $i = 1, 2$ must justify $(*)$ of Lemma 3 for $k = 0, 1, \cdots, p - 1$ and of $\sigma^p(t_1) \neq \sigma^p(t_2)$ either at least one lies in S or else they lie in the interiors of distinct intervals I_i. Notice that the finite map σ between $s_{n_k - 1}$ and s_{n_k} and the map f on I_{n_k} are either both increasing or both decreasing. Thus $\sigma^p(t_1)$ and $\sigma^p(t_2)$ have the same order (or the reversal order) as t_1 and t_2 if and only if $f^p(x_{t_1})$ and $f^p(x_{t_2})$ have the same order (or the reversed order) as x_{t_1} and x_{t_2}. But $\sigma^p(t_1)$ and $\sigma^p(t_2)$ are ordered like $f^p(x_{t_1})$ and $f^p(x_{t_2})$. Thus $x_{t_1} < x_{t_2}$.

Proof of Theorem 1. Since x_t has the same itinerary at t (Lemma 2), $f(x_t)$ must have the same itinerary as $\sigma(t)$. Lemma 3 now implies that $f(x_t) = x_{\sigma(t)}$ for $t \in T - S$. Thus $X_T = \{x_t | t \in T\}$ form a Markov portion for f which is finer than the previous partition X_S. Now suppose $f([x_{i-1}, x_i]) \supseteq [x_{j-1}, x_j]$ where we take the right-hand limit at x_{i-1}

and the left-hand limit at x_i. Then the interval with end points $x_{\sigma(i-1)}$ and $x_{\sigma(i)}$ covers $[x_{j-1}, x_j]$ under f, and $\{j-1, j\} \subseteq \{t \in T | t \text{ lies between } \sigma(i-1) \text{ and } \sigma(i)\}$, i.e., $a_{ij} = 1$. All of this reasoning is reversible and so A_σ is the adjacency matrix of f with respect to the Markov partition X_T. Now let $M_\sigma = A_\sigma$. By Lemma 1, $\rho(M_\sigma) = \rho(\lambda A_\sigma) = 1$. Therefore the spectral radius of A_σ is also given by λ. □

Remark 1. The condition that both A_μ and A_σ be irreducible in Theorem 1 is, of course, not necessary. Consider, for example, the following trivial situation:

$$A_\mu = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad A_\sigma = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

Clearly both matrices are shaped like the triangle map τ and the spectral radius of each is equal to 2. A_σ, however, is reducible since the 1 in the (3, 3) position implies the existence of an invariant subset.

Remark 2. Let τ be a piecewise linear Markov map. Let J_1 be a Markov partition under τ such that 0-1 matrix A_μ it induces is irreducible and not a permutation matrix. Let J_2 be any other finer Markov partition under τ_1 and let A_σ be the induced 0-1 matrix. Then it follows from Theorem 2.2 and Theorem 2.1 of [4] that A_σ is irreducible. Thus the irreducibility of A_σ is a necessary condition for J_2 to be a Markov partition under τ. From this we conclude that if A is not irreducible, it cannot possibly be induced by a Markov partition under τ.

Example 1. Consider the 10 × 10 matrix

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

It is easy to see that B induces the finite map, $\sigma\{0, 1, \cdots, 10\} \rightarrow \{0, 1, \cdots, 10\}$ given by the following table:

i	0	1	2	3	4	5	6	7^-	7^+	8	9	10
$\sigma(i)$	0	4	5	6	8	10	8	5	0	2	6	10

with invariant set $\{0, 5, 7^-, 7^+, 10\}$. The induced permutation μ generates the piecewise-linear Markov map shown in Fig. 4 whose adjacency matrix is

$$A_\mu = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Since $\rho(A_\mu) = 2.61803399$ and A_μ, B are irreducible, it follows from Theorem 1 that $\rho(B) = 2.61803399$.

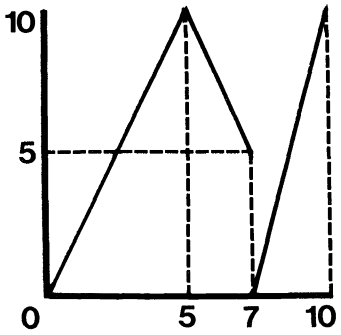


FIG. 4

Example 2. The existence of an invariant subset $S_1 \subset S$ does not in itself guarantee that the matrices associated with S_1 and S have the same spectral radii. Let

$$A_\sigma = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

be generated by the finite map σ on $S = \{0, 1, 2, 3, 4, 5, 6\}$ defined by

n	0	1	2	3	4	5	6
$\sigma(n)$	0	3	4	1	6	3	0

.

The spectral radius of A_σ is 1.56804583. Consider now the finite map μ on the invariant $S_1 = \{0, 2, 4, 6\}$ defined by

n	0	2	4	6
$\mu(n)$	0	4	6	0

where the associated matrix is

$$A_\mu = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

whose spectral radius is 2. Theorem 1 does not hold here because σ is not piecewise monotonic on S_1.

Example 3. The theorem is not true when A_σ is reducible. Consider the following:

n	0	1	2	3	4
$\sigma(n)$	1	4	3	2	0

which induces

$$A_\sigma = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

whose spectral radius is 1.914 $\cdots$ A_σ is easily seen to be reducible. Now σ has invariant set $S = \{0, 1, 4\}$ on which σ is monotonic, inducing the matrix

$$A_\mu = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

which has a different spectral radius. The question of what happens when a permutation induces a reducible matrix is considered in detail in [11].

4. Characteristic polynomials. In general, since A_μ and A_σ have the same spectral radius, they will obviously have a common factor, but there is no reason to suspect that the entire characteristic polynomial of the smaller matrix will be a factor of the characteristic polynomial of the larger matrix. The following result, however, shows that this is the case if the characteristic polynomial of the smaller matrix is irreducible with respect to the integers.

PROPOSITION 1. *Let f [0, 1] → [0, 1] be a piecewise linear Markov map of constant slope. Let $A = A_\sigma(m \times m)$ and $B = A_\mu(n \times n)$ be irreducible matrices generated by f, where $n > m$. If $C_A(x)$, the characteristic polynomial of P is irreducible with respect to the integers, then $C_A(x)$ is a factor of the characteristic polynomial of B, $C_B(x)$.*

Proof. By Theorem 1, we know that A and B have the common spectral radius ρ. By the remainder theorem,

$$C_B(x) = q(x)C_A(x) + r(x)$$

where the degree of $r(x)$ is less than the degree of $C_A(x) = m$. Since ρ is a root of both $C_A(x)$ and $C_B(x)$,

$$C_B(\rho) = q(\rho)C_A(\rho) + r(\rho)$$

implies that $r(\rho) = 0$. But ρ is a root of the irreducible polynomial $C_A(x)$ of degree m, and hence cannot be a root of a polynomial of smaller degree. Thus $r(x) = 0$.

Example 4. Let $\bar{f}$: [0, 1] → [0, 1] generate the matrices

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

B is irreducible and since the invariant set $\{0, 2, 5\}$ induces A by means of

$$\mu\colon \{0,2,5\} \to \{0,2,5\},$$

defined by $\mu(0) = 2$, $\mu(2) = 5$, $\mu(5) = 0$, Theorem 1 applies. Therefore, A and B are generated by the same map and therefore have the same spectral radius. Since $C_A(x) = x^2 - x - 1$ is an irreducible polynomial, Proposition 1 shows that this is a factor of $C_B(x)$. Examination of B reveals that the null space of B has dimension 2. Thus,

$$C_B(x) = (x - r)x^2(x^2 - x - 1)$$

where r is unknown. Let r_1, r_2 be the roots of $C_A(x)$. Since $r_1 + r_2 =$ trace of $A = 1$, and $r + r_1 + r_2 =$ trace of $B = 1$, we obtain, $r = 0$. Therefore, the dimension of the null space of B is in fact 3, and

$$C_B(x) = x^3(x^2 - x - 1).$$

Example 5. Let $\bar{f}$ be as in Example 4 and let

$$A=\begin{pmatrix}0&1&0\\0&0&1\\1&1&1\end{pmatrix},\qquad B=\begin{pmatrix}0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\\0&0&0&1&1&1\\1&1&1&0&0&0\end{pmatrix}.$$

A and B are both irreducible. From B, we obtain

$$\mu\colon \{0,1,2,3,4,5,6\}\to\{0,1,2,3,4,5,6\}$$

defined by the following table:

i	0	1	2	3	4	5	6
$\mu(i)$	2	3	4	5	6	3	0

The set $\{0, 2, 4, 6\}$ is invariant under μ and induces the matrix A. Hence Theorem 1 implies that A and B have the same spectral radius. Since $C_A(x) = x^3 - x^2 - x - 1$ is an irreducible polynomial, Proposition 1 shows that

$$C_B(x) = q(x)(x^3 - x^2 - x - 1).$$

By inspection the dimension of the null space of B is at least 1. Therefore,

$$C_B(x) = xh(x)(x^3 - x^2 - x - 1).$$

Direct computation of $C_B(x)$ yields: $h(x) = x^2 + 1$.

This example can be generalized. Consider an $n \times n$ companion matrix A, consisting of 1's on the superdiagonal and 1's in the mth row. Then,

$$C_A(x) = x^m - x^{m-1} - x^{m-2} - \cdots - x - 1.$$

It is easy to see that $C_A(x)$ has a real root $r > 1$. With the aid of Rouche's Theorem, it can be shown that all the other roots of $C_A(x)$ are inside the open unit circle. Hence r is a $P - V$ number [3, Chap. VIII]. From this it follows that $C_A(x)$ is irreducible over the integers. Therefore, if an $n \times n$ irreducible matrix B, $n > m$, can be reduced to A, they both have the same spectral radius (Theorem 1). By Proposition 1, $C_A(x)$ is a factor of $C_B(x)$.

5. Spectral radius of matrix combinations. Let $\tau \in \mathcal{C}$ induce the $n \times n$ matrix M. In [1] it is shown that M is diagonally similar to a stochastic matrix. The diagonal matrix used for this depends only on the partition $\mathcal{P}$ and not on the particular map τ. Let $\mathcal{C}_p \subset \mathcal{C}$ denote all the maps in $\mathcal{C}$ which have $\mathcal{P}$ as a partition. Thus the matrix DMD^{-1} is stochastic for all $M \in \mathcal{C}_p$. Now let $\tau, \gamma \in \mathcal{C}_p$ induce M_1 and M_2, respectively. Then, for $0 < \alpha < 1$

$$D(\alpha M_1 + (1-\alpha)M_2)D^{-1} = \alpha DM_1D^{-1} + (1-\alpha)DM_2D^{-1}$$

shows that $\alpha M_1 + (1 - \alpha)M_2$ is diagonally similar to a stochastic matrix. (Note that $\alpha M_1 + (1 - \alpha)M_2$ is not necessarily a matrix induced by some $\tau \in \mathcal{C}$.) Hence,

$$\rho(\alpha M_1 + (1-\alpha)M_2) = 1 = \alpha\rho(M_1) + (1-\alpha)\rho(M_2). \tag{1}$$

Let us assume that τ and γ have the same constant slope ± λ. Then $M_i = A_i/\lambda$, where A_i is a 0-1 matrix, $i = 1, 2$, and (1) becomes

$$\rho(\alpha A_1 + (1-\alpha)A_2) = \lambda\alpha\rho(A_1) = (1-\alpha)\lambda\rho(A_2) = \lambda. \tag{2}$$

In the special case when $\alpha = 1/2$, we get

$$\rho(A_1 + A_2) = 2\lambda. \tag{3}$$

Example 6. Consider τ and γ as shown in Fig. 5 where the common Markov partition $\mathscr{P}$ is $0 < 3/8 < 1/2 < 3/4 < 1$. Under τ and γ, $\mathscr{P}$ induces the 0-1 matrices

$$A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix},$$

respectively, where $\rho(A_1) = \rho(A_2) = \lambda = 2$. By (2), it follows that for $0 < \alpha < 1$,

$$\rho(\alpha A_1 + (1-\alpha)A_2) = 2.$$

In the special case when $\alpha = 1/2$, we get $\rho(A_1 + A_2) = 4$, where

$$A_1 + A_2 = \begin{pmatrix} 2 & 2 & 2 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 0 \end{pmatrix}.$$

The column or row estimates yields only: $2 \leqq \rho(A_1 + A_2) \leqq 6$.

Example 7. Let $\lambda = (1 + \sqrt{5})/2$ be the positive root of $x^2 - x - 1 = 0$, and consider the partition $0 = x_0 < x_1 < x_2 < x_3 < x_4 < x_5 = 1$. Let $\alpha_1 = x_1 - x_0$, $\alpha_2 = x_2 - x_1$, $\alpha_3 = x_3 - x_2$, $\alpha_4 = x_4 - x_3$, $\alpha_5 = x_5 - x_4$. Define $\alpha_1 = \alpha_5$, $\alpha_2 = \alpha_5/\lambda$, $\alpha_3 = \alpha_5$, $\alpha_4 = \alpha_5/\lambda$, where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 1$. Consider the two maps of constant slope $= \lambda$, shown in Fig. 6. Under the partition $\mathscr{P}$ defined by $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ above τ and γ induce the 0-1 matrices

$$A_1 = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

respectively. Since

$$\frac{\alpha_3 + \alpha_4}{\alpha_1} = \frac{\alpha_5}{\alpha_2} = \frac{\alpha_5 + \alpha_4}{\alpha_3} = \frac{\alpha_3}{\alpha_4} = \frac{\alpha_1 + \alpha_2}{\alpha_5} = \lambda,$$

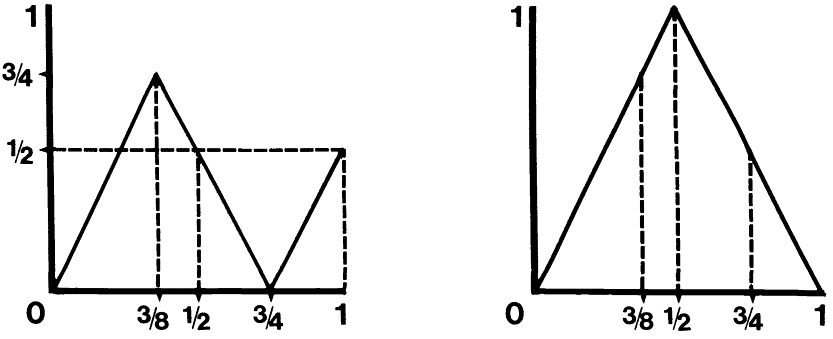


FIG. 5

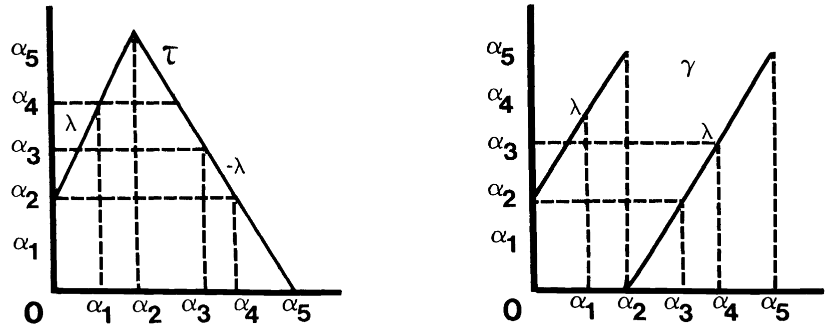


FIG. 6

$\mathcal{P}$ is a Markov partition for τ. Since $(\alpha_1 + \alpha_2)/\alpha_3 = (\alpha_4 + \alpha_5)/\alpha_5 = \lambda$, $\mathcal{P}$ is also a Markov partition for γ.

Now, both matrices A_1 and A_2 are irreducible and can be reduced to the 2×2 matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix},$$

whose spectral radius is $\rho = (1 + \sqrt{5})/2$. Thus, for example, from (3) we get that

$$A_1 + A_2 = \begin{pmatrix} 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

has spectral radius $1 + \sqrt{5}$.

REFERENCES

[1] N. FRIEDMAN AND A. BOYARSKY, *Matrices and eigenvalues induced by Markov maps*, Linear Algebra Appl., 38 (1981), pp. 141–147.

[2] W. BYERS AND A. BOYARSKY, *Absolutely continuous invariant measures that are maximal*, Trans. Amer. Math. Soc., 290 (1985), pp. 303–314.

[3] J. W. S. CASSELS, *An Introduction to Diaphontine Approximation*, Cambridge University Press, London, 1965.

[4] N. FRIEDMAN AND A. BOYARSKY, *Irreducibility and primitivity using Markov maps*, Linear Algebra Appl., 37 (1981), pp. 103–117.

[5] Z. NITECKI, *Topological Dynamics on the Interval, Ergodic Theory and Dynamical Systems* II, Birkauser, Boston, 1982, pp. 1–73.

[6] A. BOYARSKY AND M. SCAROWSKY, *On a class of transformations which have unique absolutely continuous invariant measures*, Trans. Amer. Math. Soc., 255 (1979), pp. 243–262.

[7] N. FRIEDMAN AND A. BOYARSKY, *Entropy versus speed in ergodic Markov maps*, this Journal, 5 (1984), pp. 82–93.

[8] ———, *Construction of ergodic transformations*, Adv. in Math., 43 (1982), pp. 213–254.

[9] W. PARRY, *Intrinsic Markov chains*, Trans. Amer. Math. Soc., 112 (1962), pp. 55–66.

[10] E. SENETA, *Non-negative Matrices and Markov Chains*, second ed., Springer-Verlag, Berlin-New York, 1981.

[11] W. BYERS, *Matrices induced by endomorphisms of finite sets*, Linear Algebra Appl., to appear.

© 1987 Society for Industrial and Applied Mathematics
006

AN APPLICATION OF GENERALIZED TREE PEBBLING TO SPARSE MATRIX FACTORIZATION*

JOSEPH W. H. LIU†

Abstract. A generalized version of the pebble game for trees is described. It is motivated by the study of out-of-core methods for the Cholesky factorization of sparse matrices. A solution to the generalized pebbling problem will give an equivalent ordering of the sparse matrix, so that the reordered matrix requires the minimum amount of in-core storage for its out-of-core factorization using the scheme in [12]. An efficient algorithm is presented to determine such an optimal solution.

Key words. sparse matrix, factorization, out-of-core, elimination tree, tree pebbling

AMS(MOS) subject classifications. 65F50, 65F25

1. Introduction. It is well known that many sparse matrix problems can be conveniently studied using graph-theoretic approaches. For example, the problem of reducing or minimizing bandwidth for a sparse symmetric matrix structure can be examined as a linear layout problem for graphs [4]. The fill-reduction ordering problem is closely related to the graph separator problem [10].

In this paper, we consider a problem encountered in the out-of-core solution of a sparse symmetric matrix. We want to find an equivalent ordering of a given sparse matrix, which will minimize the amount of in-core storage requirement for the successful execution of an out-of-core factorization scheme. We show that this sparse matrix problem can be transformed to a graph problem as a general form of the *pebble game* for rooted trees. This pebble game is originally introduced to study register allocation in straight-line programs [2]. It has received much attention on different variations of the basic problem [6]–[9], [14]–[16].

The generalized form of the game studied in this paper is quite different from the others in the literature. The number of pebbles required to satisfy a tree node can now be more than one. We provide an efficient algorithm to solve this generalized pebble game problem, and the underlying approach is similar to the one used by Yannakakis [20] to solve the related min-cut linear arrangement for trees. It should be noted that the algorithm can also be used to determine the best possible ordering for the out-of-core multifrontal method [3], [17] in terms of primary storage reduction.

An outline of this paper is as follows. In § 2, we describe briefly the necessary background on the sparse out-of-core factorization scheme introduced by the author in [12]. We formally introduce the class of equivalent orderings to be considered in the paper. It is based on the important tree structure, called the *elimination tree*, obtained from the sparse Cholesky factor matrix. The storage requirement on a fixed ordering for the out-of-core scheme is also derived.

In § 3, the problem of finding an optimal equivalent ordering that minimizes the primary storage requirement is transformed into the generalized pebble game problem.

* Received by the editors May 12, 1986; accepted for publication (in revised form) September 29, 1986. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A5509, by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems Inc., and by the U.S. Air Force Office of Scientific Research under contract AFOSR-ISSA-85-00083.

† Department of Computer Science, York University, North York, Ontario, Canada M3J 1P3.

A brief summary of existing pebbling algorithms to solve special forms of this pebble game is also presented.

Sections 4 to 6 are devoted to the development of an algorithm to solve the generalized pebble game. An overview of the method is given in § 4. The overall scheme makes use of the recursive structure of trees. It determines optimal orderings for subtrees, and then combines them to yield an optimal ordering for the entire tree.

Section 5 introduces the notion of a cost sequence. It is adapted from the one used by Yannakakis [20] on the min-cut layout problem for trees. This notion is essential in developing the overall optimal algorithm. Optimality is now in terms of this cost sequence together with a newly-defined partial order. In § 6, the algorithm to combine optimal subtree orderings is described. We prove that the overall ordering found is indeed optimal. The computational complexity of this algorithm is also addressed.

Section 7 contains our concluding remarks. There are three theorems in § 6, whose proofs are quite involved and lengthy. In order not to obscure the essential ideas in the paper, these proofs are postponed and presented in an appendix.

2. Statement of the problem.

2.1. Background on sparse out-of-core factorization. Let A be a given n by n sparse symmetric positive definite matrix, ordered appropriately by some fill-reducing ordering [5] (e.g., the minimum degree ordering). Let L be the (lower-triangular) Cholesky factor of A. The notations $\eta(L_{j*})$ and $\eta(L_{*j})$ are used to denote the number of nonzeros in the jth row and jth column of L, respectively.

In [12], the author proposes an out-of-core scheme for the sparse Cholesky factorization of large sparse matrices. The scheme is demonstrated to be quite effective in computing sparse Cholesky factors of extremely large matrices using auxiliary storage. It is based on the idea of matrix storage reorganization. A working storage vector in memory is provided to store nonzero entries of the Cholesky factor L. We shall refer to it as the "primary storage vector" for L.

If this primary storage can accommodate all nonzeros in the factor L, factorization will be carried out by the conventional in-core method [5]. Otherwise, this storage vector will be reorganized when need arises during the course of factorization. In each organization, only those values that are required for subsequent steps of factorization are to be retained in memory. In this way, it allows much larger problems to be solved in a given amount of primary storage, without having to rely on excessive data I/O to and from auxiliary storage. Indeed, auxiliary storage is used only to store the computed columns of the Cholesky factor.

In this out-of-core algorithm, the minimum amount of primary storage required during the computation of the jth column of L is given by

$$\sum_{k=1}^{j-1} \{\eta(L_{*k}) - \eta(L_{k*})\} + \eta(L_{*j}).$$

It is easy to see that this is actually the number of nonzero entries in the set

$$L_{[j]} = \{l_{ik} | k \leqq j \leqq i\}.$$

This rectangular window is the shaded region as illustrated in Fig. 2.1. We shall use $\eta(L_{[j]})$ to denote the number of nonzeros in this region.

Therefore, the minimum primary storage requirement for the successful completion of the entire factorization using the out-of-core scheme is

$$\max \{\eta(L_{[j]}) | 1 \leqq j \leqq n\}.$$

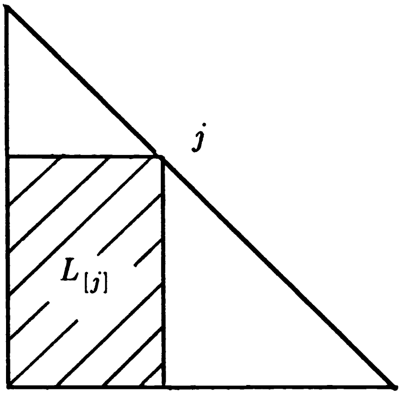


FIG. 2.1. *Region required for the computation of column j of L.*

Note that this quantity is fixed once the structure of the matrix is specified and its ordering is given.

2.2. The problem: Primary storage minimization. For a given fill-reducing ordering, it is well known that there exists a class of orderings that are equivalent in terms of fills and operations. It is based on the so-called *elimination tree* structure. This tree structure defines a class of equivalent orderings, each having the same set of filled edges as the original ordering [13], [18].

Consider the structure of the Cholesky factor L. We define the elimination tree of A to be the tree with n nodes $\{1, 2, \cdots, n\}$, and node i is the parent of node j if and only if

$$i = \min\{k | l_{kj} \neq 0\},$$

that is, i is the row subscript of the first off-diagonal nonzero in column j of L. We assume that the matrix A is *irreducible*, so that the structure is indeed a tree, and n is the root of this tree. (If A is reducible, then the elimination tree defined above is actually a forest which consists of several trees.) Figure 2.2 contains a 10-by-10 matrix example whose

$$A = \begin{pmatrix} 1 & & \mathrm{x} & & & & & & & \mathrm{x} \\ & 2 & \mathrm{x} & & & & & & & \mathrm{x} \\ \mathrm{x} & \mathrm{x} & 3 & & & & & & & \mathrm{x} \\ & & & 4 & & \mathrm{x} & \mathrm{x} & & & \\ & & & & 5 & \mathrm{x} & \mathrm{x} & & & \\ & & & \mathrm{x} & \mathrm{x} & 6 & \mathrm{x} & & & \\ & & & \mathrm{x} & \mathrm{x} & \mathrm{x} & 7 & \mathrm{x} & & \mathrm{x} \\ & & & & & & \mathrm{x} & 8 & \mathrm{x} & \mathrm{x} \\ & & & & & & & \mathrm{x} & 9 & \mathrm{x} \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & & & & \mathrm{x} & \mathrm{x} & \mathrm{x} & 10 \end{pmatrix}$$

FIG. 2.2. *A matrix example.*

diagonal entries are labeled by the corresponding equation/variable numbers. Its elimination tree is displayed in Fig. 2.3.

Any reordering that numbers nodes before parent nodes in the elimination tree is known to be equivalent to the original ordering. In other words, the number of fills and the amount of arithmetic operations to perform the factorization remain unchanged. Such orderings are referred to as *topological* orderings of the tree [19]. In this paper, we consider the problem of determining a topological ordering for a given elimination tree that will *minimize* the primary storage requirement for the out-of-core algorithm in [12].

We first re-specify the problem in graph-theoretic terms. Let $T = (X, E)$ be a given rooted tree of n nodes. For each node $x \in X$ in the tree T, two integer values are associated with it: row(x) and col(x). For any topological ordering π: $x_1, x_2, \cdots, x_n$, the *core cost* at x_j is defined to be

$$\text{core}_\pi(x_j) = \sum_{k=1}^{j-1} \{\text{col}(x_k) - \text{row}(x_k)\} + \text{col}(x_j).$$

The core cost of T with respect to the given ordering π is then

$$\max \{\text{core}_\pi(x_j) \mid 1 \leqq j \leqq n\}.$$

Our objective here is to determine an optimal ordering π that will minimize the core cost of T over all topological orderings of T.

In this paper, only topological orderings with respect to an elimination tree will be considered. Unless otherwise stated, we shall use ordering of a tree to refer to a topological ordering, that is, one that numbers nodes before parent nodes.

3. On generalized pebbling.

3.1. Problem transformation. In this section, we transform the problem in § 2 to a generalized form of a much-studied combinatorial problem: the *pebble game* for trees [8], [14], [16]. The game can be used as a model for register allocation in straight-line programs.

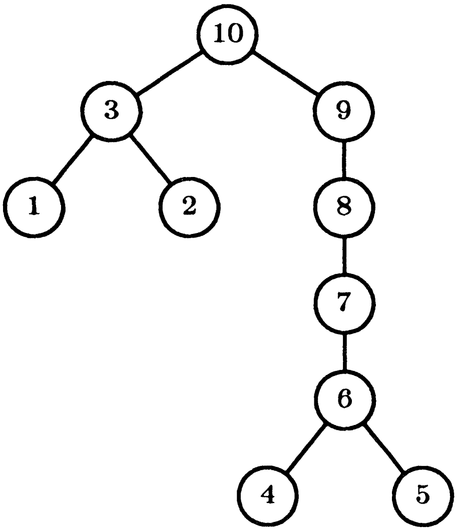


FIG. 2.3. *The elimination tree of matrix in Fig. 2.2.*

Consider a given tree T and the values row(x) and col(x) associated with each node x of T. For any (topological) ordering π: $x_1, x_2, \cdots, x_n$, the core cost as defined in the previous section can be expressed recursively as follows:

$$\text{core}_\pi(x_1) = \text{col}(x_1),$$

and for $j > 1$,

$$\text{core}_\pi(x_j) = \{\text{core}_\pi(x_{j-1}) - \text{row}(x_{j-1})\} + \text{col}(x_j).$$

It should be clear that a portion of the value $\text{core}_\pi(x_j)$ comes from nodes in the subtree rooted at x_j. This contribution from the subtree is independent of the ordering π, since nodes in the subtree under x_j are always ordered before x_j.

To aid the study of this problem, we let $T[x]$ denote the subtree of T rooted at a node x. It is also convenient to expand each original node x of T into two nodes x^+ and x^- as shown in Fig. 3.1. The node x^+ can be regarded as x during the processing of its column, with x^- as x after its processing.

Then, we can associate with each node x the two quantities:

$$\tau(x^-) = \sum_{z \in T[x]} \{\text{col}(z) - \text{row}(z)\}, \qquad \tau(x^+) = \tau(x^-) + \text{row}(x).$$

The value $\tau(x^+)$ represents the number of nonzeros in columns of L from the subtree $T[x]$, that are required *during* the processing of the column x in the factorization. On the other hand, $\tau(x^-)$ is the number of nonzeros in columns of L associated with $T[x]$ that are still required *after* the processing of x. They are the storage requirements contributed from the nodes in the subtree $T[x]$. Note that these two values depend only on the structure of the tree T, and are independent of any topological ordering.

We can now express the core cost in terms of $\tau(x^+)$ and $\tau(x^-)$. The formulation will be clearer if we introduce $\text{core}_\pi(x_j^+)$ and $\text{core}_\pi(x_j^-)$, which are the storage requirements during and after the processing of column x_j, respectively. Let

$$\text{core}_\pi(x_0^-) = 0.$$

For $j \geqq 1$, then we have

$$\text{core}_\pi(x_j^+) = \text{core}_\pi(x_{j-1}^-) + \tau(x_j^+) - \sum \{\tau(x_c^-) | x_c \text{ is a child of } x_j\},$$

$$\text{core}_\pi(x_j^-) = \text{core}_\pi(x_j^+) + \tau(x_j^-) - \tau(x_j^+).$$

This formulation actually provides a more uniform framework to study the problem. Consider the transformed tree with the $2n$ number of nodes

$$\{x_1^+, x_1^-, x_2^+, x_2^-, \cdots, x_n^+, x_n^-\}.$$

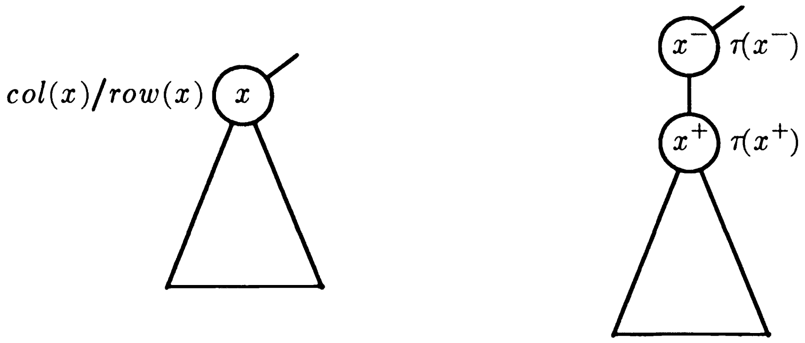


FIG. 3.1. *Tree transformation.*

Rename these nodes to $\{y_1, y_2, \cdots, y_{2n}\}$. If $\{x_j\}$ is a topological ordering on the original tree, it is clear that $\{y_j\}$ is also a topological ordering on the transformed tree. For each node, associate a τ value as follows: for $1 \leqq j \leqq n$,

$$\tau(y_{2j-1}) = \tau(x_j^+), \qquad \tau(y_{2j}) = \tau(x_j^-).$$

The transformed elimination tree of the example in Fig. 2.3 is given in Fig. 3.2. The labels in the original tree should be interpreted as "col(x)/row(x)," while that in the transformed tree are the corresponding τ-values. This tree structure will be used repeatedly throughout the paper.

Since x_j^+ is the only child node of x_j^-, the core cost in terms of the transformed tree can be collectively and conveniently expressed as:

$$\text{core}_\pi(y_0) = 0,$$

$$\text{core}_\pi(y_j) = \text{core}_\pi(y_{j-1}) + \tau(y_j) - \sum \{\tau(y_c) | y_c \text{ is a child of } y_j\}$$

for $1 \leqq j \leqq 2n$. It should be clear that an optimal topological ordering on the transformed tree in terms of the core cost will induce one on the original tree. Henceforth, we shall discard the values row(x) and col(x). Instead, we assume that a nonnegative value $\tau(y)$ is associated with each node y and core$_\pi(y)$ is defined as above in terms of $\tau(y)$.

3.2. The generalized pebble game. The transformed problem in § 3.1 can be formulated as a generalized version of the pebble game. Let T be a given rooted tree of m

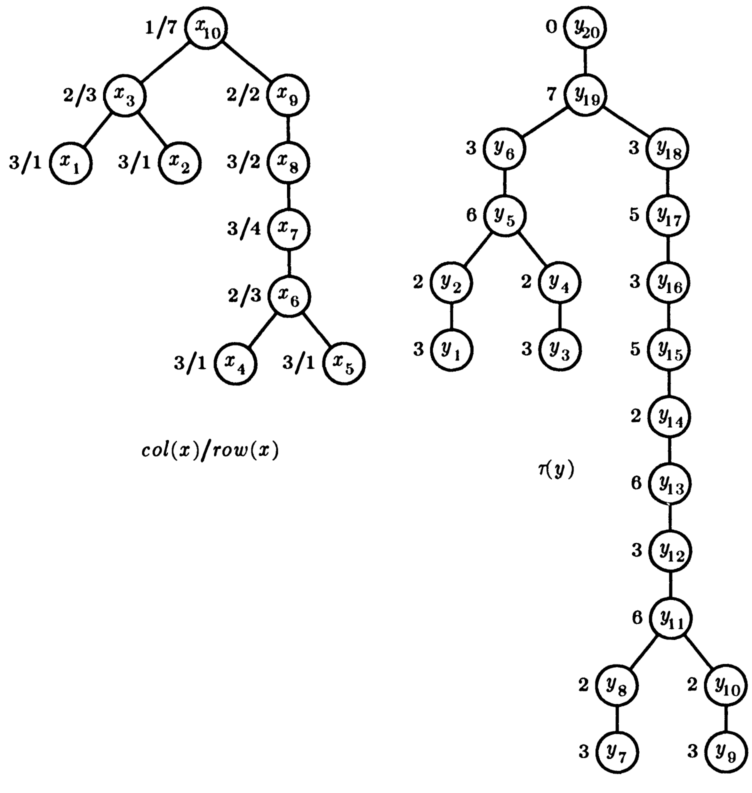


FIG. 3.2. *The transformed tree of the example in Fig.* 2.2.

nodes. For each node y in T, there is a nonnegative value $\tau(y)$ associated with it. The number $\tau(y)$ represents the number of pebbles required to satisfy the node y (a node y is said to be satisfied if there are $\tau(y)$ pebbles in this node). The generalized pebble game is played according to the following rules:

(a) If all children of an unpebbled node y are satisfied, pebbles may be placed on y (thus, a leaf node can be pebbled).

(b) If all children on an unpebbled node y are satisfied, pebbles may be moved from its children nodes to y.

(c) A pebble may be removed from a node y if there are more than $\tau(y)$ pebbles in it.

The goal of the game is, starting with no pebbles in the tree, to pebble the root of the given tree. The pebbling proceeds in moves, each move is an application of one of the above rules. The problem here is to determine a sequence of moves that will achieve the goal using the minimum number of pebbles. The sequence of moves will simply correspond to a topological ordering on the given tree.

Note that the standard (black) pebble game [8], [14] is the special case with $\tau(y) = 1$ for all nodes y in the tree. It should also be clear to the reader that the optimal solution to this generalized pebble game will be an optimal one for the primary storage minimization problem of the previous section.

3.3. Existing pebbling algorithms. The original pebble game is the special case with all pebble values $\tau(y)$ equal to 1. The solution for this standard problem can be found in [8], [14]. It is helpful to compare this scheme with the general algorithm provided later, we describe the method below. The description follows that in [8].

For a given rooted tree T with $\tau(y) = 1$ for every node y, let $p(T)$ be the minimum number of pebbles required to pebble the root. If T has only one node (the root), obviously we have $p(T) = 1$. Otherwise, assume that the root has t children nodes, and let $T_1, \cdots, T_t$ be the subtrees under the root. Then

$$p(T) = \max_{1 \leqq k \leqq t} \{p(T_k) + k - 1\},$$

where the subtrees are ordered such that

$$p(T_1) \geqq \cdots \geqq p(T_t).$$

This observation will give an algorithm that computes the value $p(T)$ and at the same time determines an (topological) ordering that achieves this minimum value. It is interesting to point out that the ordering determined by this algorithm will always number nodes within any subtree of T consecutively.

In [11], the author considers the primary storage minimization of the out-of-core multifrontal method due to Duff and Reid [3], [17]. That problem can be formulated again as a tree pebble game, where the values $\tau(y)$ can now be greater than one. However, due to the nature of the multifrontal method, *postorderings* are to be considered, that is, subtree nodes should be ordered consecutively [1]. This, therefore, may be regarded as the generalized pebble game as described in § 3.2, except for the more restrictive nature of the move sequence (postorderings). A solution to this problem is also provided in [11]. We include a brief description here for future comparison.

For a given rooted tree T with $\tau(*)$ values, let $\tilde{p}(T)$ be the minimum number of pebbles required to pebble the root, subject to the restriction that subtree nodes are to be pebbled consecutively. Assume that y is the root of T with t children: $s_1, \cdots, s_t$. Let $T_1, \cdots, T_t$ be the subtrees rooted at these children nodes.

If T has only one node y (that is, $t = 0$), $\tilde{p}(T) = \tau(y)$. Otherwise, we have

$$\tilde{p}(T) = \max\left\{\max_{1 \leqq k \leqq t}\left\{\tilde{p}(T_k) + \sum_{j=1}^{k-1} \tau(s_j)\right\}, \tau(y)\right\},$$

where the subtrees are ordered such that

$$\tilde{p}(T_1) - \tau(s_1) \geqq \cdots \geqq \tilde{p}(T_t) - \tau(s_t).$$

An algorithm, based on this observation, can be formulated to compute the value $\tilde{p}(T)$ and to determine a postordering that achieves this minimum value.

It is interesting to point out that if postorderings are not required in the out-of-core multifrontal method, the problem becomes more involved. Indeed, the algorithm to be developed in this paper will be applicable in such setting. It will give the best possible topological ordering (not necessarily postordering) so that the ordered matrix will require the least amount of primary storage in its out-of-core multifrontal factorization. For clarity, the author will focus only on the use of the ordering algorithm for the out-of-core factorization method described in § 2. Its use in the context of multifrontal method will be left to the reader.

4. Overview of strategy for optimal ordering. Given a rooted tree T of m nodes, each node y having a pebble value $\tau(y)$. Our objective is to determine a topological ordering of the tree so that the pebble game following this ordering requires the least number of pebbles.

For any (topological) ordering π: $y_1, y_2, \cdots, y_m$, define the sequence of values peb($*$):

$$\text{peb}_\pi(y_0) = 0,$$

$$\text{peb}_\pi(y_j) = \text{peb}_\pi(y_{j-1}) + \tau(y_j) - \sum\{\tau(y_c) \mid y_c \text{ is a child of } y_j\},$$

for $1 \leqq j \leqq m$. The value $\text{peb}_\pi(y_j)$ represents the total number of pebbles used during the pebbling of the node y_j; it may be appropriately called the *accumulated pebble value* at the node y_j using ordering π. The number of pebbles required to pebble the entire tree T using this ordering is given by:

$$\text{peb}_\pi(T) = \max\{\text{peb}_\pi(y_j) \mid 1 \leqq j \leqq m\}.$$

In other words, our objective is to find one such topological ordering that will minimize this pebble requirement $\text{peb}_\pi(T)$.

The recursive structure of trees can often be used to design efficient algorithms to solve problems on rooted trees. The approach is to proceed bottom up in the rooted tree. For every node y with children nodes $s_1, \cdots, s_t$, solutions are determined for all the subtrees rooted at s_k ($1 \leqq k \leqq t$). These solutions are then combined to produce one for the subtree rooted at y. A recursive use of this will solve the given problem on the overall rooted tree. Solutions to our pebbling problem are topological orderings that minimize the number of pebbles. We shall use this bottom up approach to combine optimal subtree orderings.

Let us first introduce some relevant terminology for tree orderings. Consider any rooted subtree of T, say $T[y]$, rooted at the node y. Let π be an ordering on T. The restriction of this ordering π on $T[y]$ is itself an ordering for this subtree. We shall denote this subtree ordering by $\pi[y]$ and refer to it as the *induced ordering* of π on $T[y]$. On the other hand, let ψ be an ordering on the subtree $T[y]$. ψ is said to be *compatible* with π if ψ is the same as the induced ordering $\pi[y]$.

For example, in Fig. 3.2, the induced ordering $\pi[y_{11}]$ on the subtree $T[y_{11}]$ is given by the node sequence:

$$y_7, y_8, y_9, y_{10}, y_{11}.$$

However, the following ordering on $T[y_{11}]$:

$$y_9, y_7, y_{10}, y_8, y_{11}$$

is not compatible with the original tree ordering.

Using this bottom up approach to our pebble minimization problem, we can describe our strategy as follows. Here, y is the input node with children nodes $s_1, \cdots, s_t$; and π is the returned optimal ordering for the tree $T[y]$ rooted at y.

```
ALGORITHM 4.1. Pebble-Ordering ($T[y]$, $\pi$).
    begin
      If $t = 0$ then
        return the sequence $\pi$: $y$
      else
        begin
          For $k := 1$ to $t$ do
            Pebble-Ordering ($T[s_k]$, $\psi_k$);
          Combine the optimal subtree orderings $\psi_k$, $k = 1, \cdots, t$
              to give an optimal ordering $\pi$ for $T[y]$ such that
              $\pi$ is compatible with each $\psi_k$;
        end;
    end.
```

Therefore, a strategy for optimal ordering can be obtained if we can provide an efficient solution to the *one-level* problem: combining optimal orderings of subtrees to form one for the tree. Each subtree ordering ψ_k is optimal, that is, it minimizes the value of $\text{peb}_{\psi_k}(T[s_k])$. However, this condition is not sufficient to guarantee the existence of an optimal ordering for $T[y]$ compatible with each one of the subtree orderings ψ_k.

A simple example is provided in Fig. 4.1 to illustrate this point. The ordering z_1, z_2, z_3, z_4, z_5, z_6 minimizes the pebble cost of 10 on the subtree $T[z_6]$. It is easy to verify that for all orderings on the entire tree compatible with this subtree ordering on $T[z_6]$, the pebble cost will be *at least* 14. Yet, the following ordering

$$z_1, z_2, z_4, z_5, z_7, z_8, z_3, z_6, z_9$$

will have a pebble cost of only 10.

In the next section, we introduce a new criterion for optimal orderings. We shall show that with this more involved criterion, there *always* exists an optimal compatible ordering for $T[y]$.

5. Pebble cost sequence and partial order.

5.1. Definition of pebble cost sequence. Let T be a given rooted tree of m nodes. Our objective is to find an optimal ordering $\bar{\pi}$ that minimizes the overall pebble cost in the generalized pebble game:

$$\text{peb}_{\bar{\pi}}(T) = \min \{\text{peb}_{\pi}(T) | \pi \text{ is a topological ordering}\}.$$

As noted in § 4, it is not sufficient to combine subtree orderings that minimize only the pebble costs of the subtrees. We need a more elaborate pebble cost function. This function

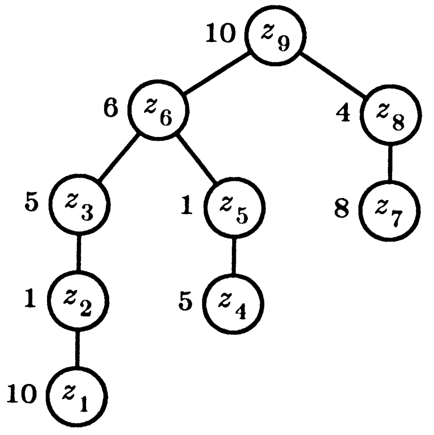


FIG. 4.1. *Example to show compatible subtree ordering.*

is adapted from the one used by Yannakakis [20] in his polynomial algorithm for the related problem of min-cut linear arrangement for trees.

Consider a topological ordering for the given tree T:

$$\pi: y_1, y_2, \cdots, y_m.$$

This defines the following sequence of values:

$$\text{peb}_\pi(y_1), \text{peb}_\pi(y_2), \cdots, \text{peb}_\pi(y_m).$$

We now introduce the *pebble cost* sequence/function. Put $v_0 = 0$. Let h_1 be the *largest* subscript of the y's such that

$$H_1 = \text{peb}_\pi(y_{h_1}) = \max \{\text{peb}_\pi(y_j) | v_0 < j \leqq m\},$$

and v_1 be the *largest* subscript such that

$$V_1 = \text{peb}_\pi(y_{v_1}) = \min \{\text{peb}_\pi(y_j) | h_1 \leqq j \leqq m\}.$$

We then define recursively h_i, v_i, and H_i, V_i as follows: h_i is the largest subscript where the *maximum* pebble cost value H_i occurs from v_{i-1} to m, and v_i the largest subscript where the *minimum* pebble cost value V_i occurs from h_i to m. Thus, we have a cost sequence, denoted by Pcost(T, π):

$$(H_1, V_1, H_2, V_2, \cdots, H_r, V_r)$$

and these values occur at the following sequence of nodes:

$$y_{h_1}, y_{v_1}, y_{h_2}, y_{v_2}, \cdots, y_{h_r}, y_{v_r}.$$

Since the tree is rooted at y_m, the last value V_r must occur at this node, that is, $v_r = m$ or $y_{v_r} = y_m$. Note also that the value of r depends on the tree structure, the pebble values and the ordering.

To illustrate the notion of this cost sequence, we consider the example in Fig. 3.2. It is clear that the pebble cost sequence for the tree is given by:

$$\text{Pcost}(T, \pi) = (9, 0),$$

and they occur at the nodes:

$$(y_{13}, y_{20}).$$

However, if we only consider the subtree $T[y_{18}]$ in this example with the induced ordering $\pi[y_{18}]$, the pebble cost sequence is then:

$$\text{Pcost}(T[y_{18}], \pi[y_{18}]) = (6, 2, 5, 3),$$

and these values occur at the nodes:

$$(y_{13}, y_{14}, y_{17}, y_{18}).$$

Note that the pebble requirement in the subtree $T[y_6]$ does not affect the pebble sequence for $T[y_{18}]$.

We shall sometimes refer to the locations y_{h_i} as the *hills* and y_{v_i} as the *valleys* of the given tree and ordering. The quantities H_i and V_i are also referred to as the *hill* and *valley* values, respectively. The motivation for the choice of these terminologies should be clear from the plot of accumulated pebble cost values $\text{peb}_\pi(y_j)$ against y_j. The plot for the subtree $T[y_{18}]$ of Fig. 3.2 is illustrated in Fig. 5.1.

Let $\text{Pcost}(T, \pi) = (H_1, V_1, \cdots, H_r, V_r)$ be a cost sequence. It is clear from definition that

$$H_1 = \text{peb}_\pi(T).$$

The following property is also obvious.

LEMMA 5.1. $H_1 > H_2 > \cdots > H_r \geqq V_r > \cdots > V_1 > V_0 = 0$.

5.2. A partial order for pebble cost sequences. We want to compare different topological orderings on a rooted tree with respect to their pebble cost sequences. To prepare for that, we introduce a *partial order* on these sequences. Let α and β be two pebble cost sequences:

$$\alpha = (\tilde{H}_1, \tilde{V}_1, \cdots, \tilde{H}_{\tilde{r}}, \tilde{V}_{\tilde{r}}), \qquad \beta = (H_1, V_1, \cdots, H_r, V_r).$$

We say that $\alpha \prec \beta$ if and only if for every i $(1 \leqq i \leqq \tilde{r})$, there exists a j $(1 \leqq j \leqq r)$ such that

$$\tilde{H}_i \leqq H_j \quad \text{and} \quad \tilde{V}_i \leqq V_j.$$

THEOREM 5.2. *"$\prec$" is a partial order on cost sequences.*

Proof. It is obvious that "$\prec$" is transitive and reflexive. It remains to show that it is anti-symmetric. Let

$$\alpha = (\tilde{H}_1, \tilde{V}_1, \cdots, \tilde{H}_{\tilde{r}}, \tilde{V}_{\tilde{r}}),$$
$$\beta = (H_1, V_1, \cdots, H_r, V_r).$$

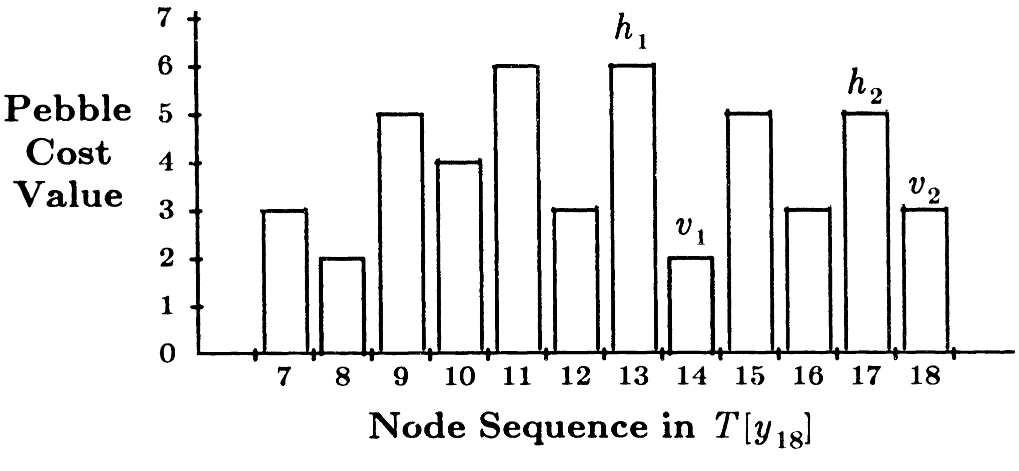


FIG. 5.1. *Plot of pebble cost for subtree* $T[y_{18}]$ *in Fig.* 3.2.

Assume that $\alpha \prec \beta$ and $\beta \prec \alpha$. Consider any $\tilde{H}_i$ and $\tilde{V}_i$. By the definition of "$\prec$", there exists a j such that $\tilde{H}_i \leqq H_j$ and $\tilde{V}_i \leqq V_j$. Since $\beta \prec \alpha$, for this j, there is a k such that $H_j \leqq \tilde{H}_k$ and $V_j \leqq \tilde{V}_k$. Combining we have $\tilde{H}_i \leqq \tilde{H}_k$ and $\tilde{V}_i \leqq \tilde{V}_k$ so that by Lemma 5.1, we must have $i = k$. This implies that $\tilde{H}_i = H_j$ and $\tilde{V}_i = V_j$.

It remains to show that for every $1 \leqq i \leqq \tilde{r}$,

$$\tilde{H}_i = H_i \quad \text{and} \quad \tilde{V}_i = V_i.$$

We prove this by induction on i. For $i = 1$, by the property established above, there exists a j such that $\tilde{H}_1 = H_j$. Assume for contradiction that $j \neq 1$, so that by Lemma 5.1 $H_j < H_1$. Then since $\beta \prec \alpha$, for H_1, there must be a k such that $H_1 \leqq \tilde{H}_k$. Combining, we have

$$\tilde{H}_1 \leqq H_j < H_1 \leqq \tilde{H}_k.$$

This contradicts Lemma 5.1 on the cost sequence α. Therefore, $\tilde{H}_1 = H_1$ (so that $\tilde{V}_1 = V_1$).

The same argument can be used for the inductive step. Therefore the sequence α must be an initial subsequence of β. By symmetry, β must also be an initial subsequence of α. Hence, α and β must be identical cost sequences. □

The next theorem follows directly from definition. It shows the relevance of the pebble cost sequence and the partial order "$\prec$" in the context of pebble minimization.

THEOREM 5.3. *For two orderings ψ and π of the tree T, if*

$$\text{Pcost}(T, \psi) \prec \text{Pcost}(T, \pi)$$

then $\text{peb}_\psi(T) \leqq \text{peb}_\pi(T)$.

The implication of this simple observation is that in order to determine an optimal ordering that minimizes the overall pebble requirement, we can restrict our search for an ordering $\bar{\pi}$ (if it exists) such that

$$\text{Pcost}(T, \bar{\pi}) \prec \text{Pcost}(T, \pi)$$

for all orderings π.

6. Combining subtree orderings.

6.1. Combine algorithm based on subtree segments. In this section, we show how to solve the *one-level* problem: combining optimal subtree orderings to give an optimal ordering for the overall tree. Here, optimality is with reference to the pebble cost sequence and the partial order "$\prec$" introduced in the last section.

Let T be a given tree rooted at the node y, and $s_1, s_2, \cdots, s_t$ be the children nodes of y. Assume that $\psi_1, \cdots, \psi_t$ are given orderings on the respective subtrees $T[s_1], \cdots, T[s_t]$.

We want to construct an optimal ordering for T which is compatible with each subtree ordering. Obviously, the last node in this ordering must be y, the root. The problem is how to interleave nodes from the t subtrees under y so that the resulting pebble cost sequence is minimized. The idea is quite simple: for each hill value in a subtree, we should try to use appropriately-chosen valley values for the remaining subtrees. This will help to reduce the impact of the hill value on the pebble cost sequence.

To facilitate the discussion, we introduce the notion of *valley segments* for an ordered tree. Consider a subtree $T[y]$ with an ordering ψ. Let its pebble cost sequence be:

$$\text{Pcost}(T[y], \psi) = (H_1, V_1, \cdots, H_r, V_r),$$

and let these values occur at the nodes

$$y_{h_1}, y_{v_1}, \cdots, y_{h_r}, y_{v_r}.$$

There are r valley segments of $T[y]$; for $1 \leqq k \leqq r$, the kth valley segment consists of the nodes

$$y_{v_{k-1}+1}, y_{v_{k-1}+2}, \cdots, y_{v_k}.$$

In other words, it is the sequence of nodes in between two valley nodes (including y_{v_k} but not $y_{v_{k-1}}$). We shall define its *segment value* to be $H_k - V_k$.

For example, consider the subtree $T[y_{18}]$ in Fig. 3.2. There are two valley segments:

$$y_7, y_8, y_9, y_{10}, y_{11}, y_{12}, y_{13}, y_{14}, \qquad y_{15}, y_{16}, y_{17}, y_{18}$$

and their segment values are 4 and 2, respectively. But the subtree $T[y_6]$ has only one segment:

$$y_1, y_2, y_3, y_4, y_5, y_6$$

which is the entire subtree, and its segment value is 3.

Valley nodes are appropriate locations to switch from one subtree to another when combining subtree orderings. Valley segments are relevant notions, and nodes within each segment can be treated as an entity. Indeed, the following algorithm combines the given subtree orderings based on an arrangement of the segments in all subtrees. As before, ψ_k is a subtree ordering of $T[s_k]$, where $s_1, \cdots, s_t$ are children nodes of the root y in the tree T.

```
ALGORITHM 6.1. Combine ($T[y]$, $\psi$)
    begin
      For $k := 1$ to $t$ do
        Determine the valley segments of the subtree $T[s_k]$
          using the cost sequence Pcost($T[s_k]$, $\psi_k$);
      Arrange the segments from all the subtrees in descending order of their
        segment values: {hill value-valley value};
      Based on this segment arrangement, order the nodes in each segment
        consecutively, followed by the root $y$;
      Return this ordering as $\psi$
    end.
```

We shall use the notation $\Phi(\psi_1, \psi_2, \cdots, \psi_t)$ to refer to the ordering ψ on $T[y]$ obtained by Algorithm 6.1. When $t = 1$, this ordering can be obtained simply by appending the root y to the subtree ordering ψ_1 of its only subtree.

It is easy to see that the ordering obtained by Algorithm 6.1 is compatible with each subtree ordering ψ_k. Indeed, the segments within each subtree are already in descending sequence with respect to their segment values (it follows from Lemma 5.1). This means the relative order of nodes in each subtree is always preserved by the new ordering.

On applying Algorithm 6.1 to the subtree $T[y_{19}]$ of the example in Fig. 3.1, we note that the root y_{19} has two children nodes y_6 and y_{18}. The subtree $T[y_6]$ has one segment of value 3; while the subtree $T[y_{18}]$ has two segments of value 4 and 2, respectively. Therefore the ordering returned by Algorithm 6.1 will be the nodes in the segment (with value 4):

$$y_7, y_8, y_9, y_{10}, y_{11}, y_{12}, y_{13}, y_{14},$$

followed by the segment (with value 3):

$$y_1, y_2, y_3, y_4, y_5, y_6,$$

then by (with value 2):

$$y_{15}, y_{16}, y_{17}, y_{18},$$

and finally by the node y_{19}. With this new ordering, the pebble cost sequence for T is reduced from (9, 0) to (8, 0).

6.2. Properties of the combine algorithm. We shall state some important properties of the ordering $\Phi(\psi_1, \cdots, \psi_t)$ obtained from Algorithm 6.1. The detailed proofs are lengthy, and we shall provide them at the end of the paper in the Appendix. The following sequence of theorems is to establish the optimality of the "Combine" algorithm when used recursively in the "Pebble-Ordering" algorithm of § 4.

THEOREM 6.1. *Let $\psi = \Phi(\psi_1, \cdots, \psi_t)$. For any ordering π' that orders nodes within each subtree segment consecutively and is compatible with each ψ_k,*

$$\mathrm{Pcost}(T, \psi) \prec \mathrm{Pcost}(T, \pi').$$

THEOREM 6.2. *Let π be any topological ordering on the tree $T[y]$, which is compatible with each subtree ordering ψ_k. There exists an ordering π' on $T[y]$, that orders nodes in subtree segments consecutively, such that*

$$\mathrm{Pcost}(T, \pi') \prec \mathrm{Pcost}(T, \pi).$$

THEOREM 6.3. *Let $\bar{\psi}_k$ be another subtree ordering for $T[s_k]$, where*

$$\mathrm{Pcost}(T[s_k], \bar{\psi}_k) \prec \mathrm{Pcost}(T[s_k], \psi_k).$$

If $\bar{\pi} = \Phi(\psi_1, \cdots, \bar{\psi}_k, \cdots, \psi_t)$, and $\psi = \Phi(\psi_1, \cdots, \psi_k, \cdots, \psi_t)$, then

$$\mathrm{Pcost}(T, \bar{\pi}) \prec \mathrm{Pcost}(T, \psi).$$

The proofs of Theorems 6.1–6.3 are left to the Appendix. Theorem 6.1 says that the cost sequence returned from Algorithm 6.1 is the smallest possible (in terms of "$\prec$") among all orderings that are based on the valley segments. Theorem 6.2 implies that if it is the smallest among segment-based orderings, it will also be the smallest among all orderings compatible with the individual subtree orderings. Finally, Theorem 6.3 points out the effect of an improved subtree ordering on the combined ordering Φ. We can now use these results to establish the optimality of our overall ordering algorithm.

THEOREM 6.4. *Let $\bar{\pi}$ be the ordering on $T[y]$ returned from Algorithm 4.1 ("Pebble-Ordering"), where subtree orderings are combined by Algorithm 6.1 ("Combine"). Then for any topological ordering π of $T[y]$,*

$$\mathrm{Pcost}(T, \bar{\pi}) \prec \mathrm{Pcost}(T, \pi).$$

Proof. We prove the result by induction on the number m of nodes in the tree $T[y]$. The result is obviously true if $m = 1$. Assume that the result is true for all trees with less than m modes. Let the children nodes of y be $s_1, \cdots, s_t$; and $\bar{\psi}_k$ be the ordering obtained from the execution of "Pebble-Ordering ($T[s_k]$, $\bar{\psi}_k$)." So $\bar{\pi}$ can be expressed as $\Phi(\bar{\psi}_1, \cdots, \bar{\psi}_t)$.

Consider any ordering π of $T[y]$, and their induced subtree orderings $\pi[s_k]$, for $1 \leqq k \leqq t$. Let π' be the ordering $\Phi(\pi[s_1], \cdots, \pi[s_t])$. By Theorems 6.1 and 6.2, π' has the best pebble cost sequence relative to all orderings compatible with each subtree ordering $\pi[s_k]$. In other words,

$$\mathrm{Pcost}(T, \pi') \prec \mathrm{Pcost}(T, \pi).$$

But, by the inductive assumption, in each subtree $T[s_k]$,

$$\mathrm{Pcost}(T[s_k], \bar{\psi}_k) \prec \mathrm{Pcost}(T[s_k], \pi[s_k]).$$

A repeated application of Theorem 6.3 and the transitive property of the partial order "$\prec$" (Theorem 5.2) will give

$$\mathrm{Pcost}(T, \bar{\pi}) \prec \mathrm{Pcost}(T, \pi').$$

Therefore, the result follows. □

Theorem 6.4 shows that Algorithms 4.1 and 6.1 can be used to yield a topological ordering $\bar{\pi}$ that minimizes the pebble cost sequence Pcost(T, $\bar{\pi}$) and hence the pebble cost value $\mathrm{peb}_{\bar{\pi}}(T)$. We now determine the time complexity of this algorithm. We show that Algorithm 4.1 (Pebble-Ordering) and Algorithm 6.1 (Combine) can be implemented in time $O(m^2)$, where m is the number of nodes in the tree.

Assume that the given tree $T[y]$ is rooted at y with m nodes, and the root y has t children nodes. Since the valley segments within each subtree are already in descending sequence with respect to their segment values, we need only to *merge* the segments from the t subtrees. This can be implemented efficiently by the multiway merge [1], and it will take at most $\{m \log_2 t\}$ time units to perform the t-way merge. Furthermore, the computation of the new pebble cost sequence on the tree requires at most m time units. Therefore, if $f(m)$ is the amount of work to execute Algorithm 4.1 using Algorithm 6.1 for combining subtree orderings, then

$$f(m) = m \log_2 t + m + \sum_{1 \leqq k \leqq t} f(m_k),$$

where m_k is the number of nodes in the kth subtree under the node y. This means that $\sum_k m_k = m - 1$.

A simple induction on m will show that $f(m) \leqq m^2$. This upper bound, though attainable, is often too pessimistic. In practice, the number of hill/valley values in the pebble cost sequence Pcost is often much smaller than the number of nodes in the subtree, so that the amount of work required for the merging of segments is usually much smaller than $\{m \log_2 t\}$. Indeed, in the application of this algorithm for storage minimization for the out-of-core sparse matrix factorization, the execution time will usually be linear with respect to the order of the matrix.

7. Concluding remarks. We have shown that the core storage minimization problem for the out-of-core factorization scheme in [12] can be studied using a generalized form of the combinatorial problem of pebble game. An efficient algorithm is provided to solve this generalized pebble game problem. It is based on the notion of cost sequences, adapted from Yannakakis [20].

It is interesting to compare the algorithm provided in § 6 with the two existing algorithms in § 3 for solving the standard pebble game and for solving the general game by postorderings. In the case of the standard pebble game where each pebble value is 1, the cost sequence of each subtree is of the form:

$$(H, 1) = (p(T), 1),$$

where $H = p(T)$ is the hill value for the subtree, and 1 is (necessarily) the valley value at the root of the subtree. Ordering the subtrees in descending sequence of the subtree hill values $\{p(T_k)\}$ is obviously equivalent to ordering them in descending sequence of the subtree segment values $\{p(T_k) - 1\}$. Therefore, the algorithm in § 3 for this standard game is a special case of the general algorithm in § 4.

On the other hand, the use of postorderings implies the use of a restricted form of the cost sequence. The restricted cost sequence can be taken to be of the form:

$$(H, V) = (\tilde{p}(T), \tau(y)),$$

where $H = \tilde{p}(T)$ is the first hill value for the subtree T and $V = \tau(y)$ is the pebble value for the root y of this subtree. Indeed, the algorithm described in § 3 can be viewed as one that orders the subtrees in descending sequence of their segment values $\{\tilde{p}(T_k) - \tau(s_k)\}$ (except that there is only one segment for each subtree).

The methodology provided in this paper to solve the generalized tree pebble game should be of theoretical and algorithmic interest. Currently, in the out-of-core sparse factorization scheme of [12], postorderings are used. In practice, it is simple to implement, and is demonstrated to be very effective. Although one can construct matrix structures to show that postorderings are not sufficient in general for primary storage *minimization*, it should still be highly recommended. More practical justification seems to be warranted for the use of the optimal algorithm presented in this paper in the context of out-of-core factorization.

Appendix.

A.1. Best subtree segment arrangement (Theorem 6.1). In this appendix, we provide detailed proofs for Theorems 6.1–6.3. We first establish the following lemma which is useful to compare two cost sequences based on the partial order "$\prec$."

LEMMA A.1. *Let* π: $y_1, y_2, \cdots, y_m$ *and*

$$\text{Pcost}(T, \pi) = (H_1, V_1, \cdots, H_r, V_r).$$

For two values $\tilde{H}$ *and* $\tilde{V}$, $\tilde{H} \leqq H_j$ *and* $\tilde{V} \leqq V_j$ *for some* j *if and only if there exists some node* y_q *in the sequence* π *such that*

$$\tilde{H} \leqq \text{peb}_\pi(y_q), \qquad \tilde{V} \leqq \min\,\{\text{peb}_\pi(y_p) | q \leqq p \leqq m\}.$$

Proof. Let $(y_{h_1}, y_{v_1}, \cdots, y_{h_r}, y_{v_r})$ be the nodes at which the values of the pebble cost sequence Pcost(T, π) occur.

"*if part.*" Let the node y_q in the lemma be in the segment between the valley nodes $y_{v_{j-1}}$ and y_{v_j}. From definition, we have

$$\tilde{H} \leqq \text{peb}_\pi(y_q) \leqq \text{peb}_\pi(y_{h_j}) = H_j.$$

Moreover, $q \leqq v_j$, so that by the condition on $\tilde{V}$ in this lemma,

$$\tilde{V} \leqq \text{peb}_\pi(y_{v_j}) = V_j.$$

"*only if part.*" Let $\tilde{H} \leqq H_j$ and $\tilde{V} \leqq V_j$. Then take $q = h_j$. The result is obvious. □

Let us follow the same notations as in § 6.1. That is, let the given tree T be rooted at the node y, which has $s_1, \cdots, s_t$ as its children nodes. To help the discussions and formal proofs, we first introduce a definition.

Consider a node z in the tree. The pebble cost of z in T with ordering π is given by $\text{peb}_\pi(z)$. We shall use the notation $\text{peb}_{\pi[s_k]}(z)$ to denote the pebble contribution to the value $\text{peb}_\pi(z)$ from nodes in the subtree $T[s_k]$. Some properties of this value are expressed in the next lemma, and the proofs are straightforward and are omitted.

LEMMA A.2. (a) $\text{peb}_\pi(z) = \sum_{k=1}^{t} \text{peb}_{\pi[s_k]}(z)$.

(b) *If the node* z *belongs to the subtree* $T[s_k]$, *then* $\text{peb}_{\pi[s_k]}(z)$ *is simply the pebble cost at the node* z *of the tree* $T[s_k]$ *using the induced ordering* $\pi[s_k]$.

(c) *If the node* z *does not belong to the subtree* $T[s_k]$, *then*

$$\text{peb}_{\pi[s_k]}(z) = \text{peb}_{\pi[s_k]}(x_k)$$

where x_k *is the last node from the subtree* $T[s_k]$ *appearing before* z *in the sequence* π. □

COROLLARY A.3. *Consider two orderings* $\tilde{\pi}$ *and* π *on the tree* T *such that* $\tilde{\pi}[s_k] = \pi[s_k]$, *that is, the same subtree ordering when restricted to* $T[s_k]$. *For two nodes*

$\tilde{z}$ and z, let $\tilde{x}_k$ and x_k be the last node from $T[s_k]$ appearing before and including $\tilde{z}$ and z in the ordering $\tilde{\pi}$ and π, respectively:

$$\tilde{\pi}: \cdots \cdots \tilde{x}_k \cdots \tilde{z} \cdots, \qquad \pi: \cdots \cdots x_k \cdots z \cdots,$$

(a) *if* $\tilde{x}_k = x_k$, *then* $\text{peb}_{\tilde{\pi}[s_k]}(\tilde{z}) = \text{peb}_{\pi[s_k]}(z)$,

(b) *if* $\text{peb}_{\tilde{\pi}[s_k]}(\tilde{x}_k) \leqq \text{peb}_{\pi[s_k]}(x_k)$, *then* $\text{peb}_{\tilde{\pi}[s_k]}(\tilde{z}) \leqq \text{peb}_{\pi[s_k]}(z)$. □

Note that in Corollary A.3, $\tilde{\pi}$ and π can be the same ordering. To illustrate the results, consider the ordering in Fig. 3.2 and the two subtrees under the node y_{19} with children $s_1 = y_6$ and $s_2 = y_{18}$, we have

$$\text{peb}_\pi(y_4) = \text{peb}_{\pi[y_6]}(y_4) + \text{peb}_{\pi[y_{18}]}(y_4) = 4 + 0 = 4,$$

$$\text{peb}_\pi(y_9) = \text{peb}_{\pi[y_6]}(y_9) + \text{peb}_{\pi[y_{18}]}(y_9) = 3 + 5 = 8.$$

Note also that

$$\text{peb}_{\pi[y_6]}(y_j) = 3 \quad \text{for all } 7 \leqq j \leqq 18,$$

$$\text{peb}_{\pi[y_{18}]}(y_j) = 0 \quad \text{for all } 1 \leqq j \leqq 6.$$

We are now ready to examine properties of orderings that are based on subtree segments, that is, nodes in each subtree segment are ordered consecutively. As in § 6.1, let $\psi_1, \psi_2, \cdots, \psi_t$ be given orderings on the respective subtrees $T[s_1], T[s_2], \cdots, T[s_t]$. These orderings on the subtrees define segments based on their individual valley values. We shall use the term *segment ordering* to refer to any ordering on the entire tree that numbers nodes in each subtree segment consecutively. In other words, each segment ordering corresponds to an arrangement of the subtree segments followed by the node y. The proof of the next lemma is straightforward and is omitted.

LEMMA A.4. *Let π be a segment ordering on T that is compatible with each subtree ordering ψ_k. Let (H, V) be a hill/valley value pair in the pebble cost sequence* Pcost(T, π).

(a) *The hill value H occurs either at a hill location in some subtree $T[s_k]$ or at the node y.*

(b) *The valley value V occurs either at a valley location in some subtree $T[s_k]$ or at the node y.*

(c) *If H occurs at a hill node x in the subtree $T[s_k]$, then V occurs at the valley node in this subtree immediately following x or at the root y.* □

THEOREM A.5. *Let π be a segment ordering on T that is compatible with each subtree ordering ψ_k. Interchanging any two neighboring (subtree) segments that are not in descending sequence of their segment values will not increase the pebble cost sequence.*

Proof. Consider two neighboring (subtree) segments that are not in sequence with respect to their segment values. The two segments must come from two different subtrees, since π maintains the relative order of segments for each subtree and segments from the same subtree are already in descending sequence.

For concreteness, let the first segment belong to the subtree $T[s_a]$ with h_a and v_a as its hill and valley nodes, respectively. Also let $T[s_b]$, h_b, v_b correspond to the second segment. We can view the given ordering as:

$$\pi: \cdots(\cdots h_a \cdots v_a)(\cdots h_b \cdots v_b)\cdots$$

where parentheses are used here to identify the two segments. The given condition in the theorem can be expressed as:

$$(**) \qquad \text{peb}_{\pi[s_a]}(h_a) - \text{peb}_{\pi[s_a]}(v_a) \leqq \text{peb}_{\pi[s_b]}(h_b) - \text{peb}_{\pi[s_b]}(v_b)$$

since by Lemma A.2(b), the left- and right-hand sides are the segment values of the two subtree segments.

Now consider the new ordering $\tilde{\pi}$ by interchanging *only* these two segments:

$$\tilde{\pi}: \cdots(\cdots h_b \cdots v_b)(\cdots h_a \cdots v_a)\cdots.$$

We are to show that $\text{Pcost}(T, \tilde{\pi}) \prec \text{Pcost}(T, \pi)$. By Lemma A.1, for each hill/valley pair of the new sequence $\tilde{\pi}$, it is sufficient to find a node y_q in the sequence π satisfying the conditions in that lemma. Consider any hill/valley value pair (H, V) in the sequence $\tilde{\pi}$. Let the hill value occur at the node x.

Case 1. x is outside the two segments under consideration. Then by Lemma A.4(a), x must be either the root y or a hill location in one subtree. Choose y_q to be the same node x in the π sequence, and it is easy to verify that this node satisfy the conditions in Lemma A.1.

Case 2. x belongs to the segment $(\cdots h_b \cdots v_b)$. By Lemma A.4(a), x must be the node h_b. The node y_q for Lemma A.1 will be chosen to be h_b $(=x)$ in the π sequence. Indeed, applying Corollary A.3, we have

$$\begin{aligned}
\text{peb}_{\tilde{\pi}}(x) &= \sum_{k \neq a} \text{peb}_{\tilde{\pi}[s_k]}(h_b) + \text{peb}_{\tilde{\pi}[s_a]}(h_b)\\
&= \sum_{k \neq a} \text{peb}_{\pi[s_k]}(h_b) + \text{peb}_{\tilde{\pi}[s_a]}(h_b)\\
&\leqq \sum_{k \neq a} \text{peb}_{\pi[s_k]}(h_b) + \text{peb}_{\pi[s_a]}(v_a)\\
&= \sum_{k \neq a} \text{peb}_{\pi[s_k]}(h_b) + \text{peb}_{\pi[s_a]}(h_b) = \text{peb}_{\pi}(h_b).
\end{aligned}$$

By Lemma A.4, we have

$$V = \min\{\text{peb}_{\tilde{\pi}}(v_b), \text{peb}_{\tilde{\pi}}(y)\}.$$

Applying Corollary A.3, we have

$$\text{peb}_{\tilde{\pi}[s_a]}(v_b) \leqq \text{peb}_{\pi[s_a]}(v_b)$$

so that

$$\text{peb}_{\tilde{\pi}}(v_b) \leqq \text{peb}_{\pi}(v_b).$$

Therefore, the value V must be less than the accumulated pebble value of any node after h_b in π.

Case 3. x belongs to the segment $(\cdots h_a \cdots v_a)$. This means that $x = h_a$. We shall choose y_q for Lemma A.1 again to be the node h_b in π. Again by Corollary A.3, we have

$$\begin{aligned}
\text{peb}_{\tilde{\pi}}(x) &= \sum_{k \neq a,b} \text{peb}_{\tilde{\pi}[s_k]}(h_a) + \text{peb}_{\tilde{\pi}[s_a]}(h_a) + \text{peb}_{\tilde{\pi}[s_b]}(h_a)\\
&= \sum_{k \neq a,b} \text{peb}_{\pi[s_k]}(h_a) + \text{peb}_{\pi[s_a]}(h_a) + \text{peb}_{\tilde{\pi}[s_b]}(v_b)\\
&= \sum_{k \neq a,b} \text{peb}_{\pi[s_k]}(h_a) + \text{peb}_{\pi[s_a]}(h_a) + \text{peb}_{\pi[s_b]}(v_b).
\end{aligned}$$

But, by the given condition $(**)$ on the two segments in π, this value must be no greater than

$$\begin{aligned}
\sum_{k \neq a,b} \text{peb}_{\pi[s_k]}(h_a) + \text{peb}_{\pi[s_b]}(h_b) + \text{peb}_{\pi[s_a]}(v_a) &= \sum_{k \neq a,b} \text{peb}_{\pi[s_k]}(h_b) + \text{peb}_{\pi[s_b]}(h_b) + \text{peb}_{\pi[s_a]}(h_b)\\
&= \text{peb}_{\pi}(h_b).
\end{aligned}$$

The condition on the value V can be verified in the same way as in Case 2.

Therefore, in all cases, for a given pair (H, V) in $\text{Pcost}(T, \tilde{\pi})$, we can bound them by a corresponding pair from $\text{Pcost}(T, \pi)$. It follows from definition that

$$\text{Pcost}(T, \tilde{\pi}) \prec \text{Pcost}(T, \pi). \qquad \square$$

Proof of Theorem 6.1. Let $\psi = \Phi(\psi_1, \cdots, \psi_t)$. Consider any given segment ordering π' compatible with each subtree ordering ψ_k. A finite number of neighboring subtree segment interchanges will transform π' to ψ. Repeated applications of Theorem A.5 for such interchanges together with the transitivity of "$\prec$" (Theorem 5.2) will show that

$$\text{Pcost}(T, \psi) \prec \text{Pcost}(T, \pi'). \qquad \square$$

A.2. Segment orderings are sufficient (Theorem 6.2). Theorem 6.2 says that in order to search for an ordering that will minimize the cost sequence, it is sufficient to look for a segment ordering that is compatible with the subtree orderings. The following is a constructive proof.

Proof of Theorem 6.2. Let π be any given ordering on the tree T rooted at y, which is compatible with each subtree ordering ψ_k of the subtree $T[s_k]$. We shall prove the result by constructing a segment ordering π' such that

$$\text{Pcost}(T, \pi') \prec \text{Pcost}(T, \pi).$$

Construct the new ordering π' from π as follows:

(a) Remove all nodes from the sequence π except the root y and hill locations of the subtrees;

(b) Replace each hill location by the subtree segment associated with it.

It should be clear that π' is still compatible with each subtree ordering ψ_k, and orders nodes in each subtree segment consecutively. Moreover, it maintains the relative order of all the subtree hill nodes in the original ordering π. It remains to show that the pebble cost sequence of π' is no greater than that of π.

Consider any hill/valley value pair (H, V) in $\text{Pcost}(T, \pi')$. Let the hill value occur at the node x. It is sufficient to find a node y_q in the original sequence π satisfying the conditions in Lemma A.1. If x is the root y, pick this root as the node y_q and the conditions in Lemma A.1 are clearly satisfied. Otherwise, by Lemma A.4(a), x must be a hill node in one of the subtrees. Let $x = h_a$ belonging to the subtree $T[s_a]$, and v_a be the valley node immediately following h_a in the pebble cost sequence of this subtree. That is

$$\pi' \colon \cdots (\cdots h_a \cdots v_a) \cdots .$$

We now show that y_q for Lemma A.1 can be chosen to be the node x. We first claim that for each k, $\text{peb}_{\pi'[s_k]}(x) \leqq \text{peb}_{\pi[s_k]}(x)$. By Corollary A.3, since x belongs to $T[s_a]$, we have

$$\text{peb}_{\pi'[s_a]}(x) = \text{peb}_{\pi[s_a]}(x).$$

For $k \neq a$, let the last segment from the subtree $T[s_k]$ before x in π' be $(\cdots h_k \cdots v_k)$. (If no such segment exists, then $\text{peb}_{\pi'[s_k]}(x) = 0$, and the result holds.) By Corollary A.3,

$$\text{peb}_{\pi'[s_a]}(x) = \text{peb}_{\pi'[s_a]}(v_k) = \text{peb}_{\pi[s_a]}(v_k).$$

Let x_k be the last node from $T[s_k]$ before x in π. Since π' maintains the relative order of all the subtree hill values, x_k must appear after the hill node h_k of $T[s_k]$. By the definition of the valley node v_k and Corollary A.3, we have

$$\text{peb}_{\pi[s_k]}(v_k) \leqq \text{peb}_{\pi[s_k]}(x_k) = \text{peb}_{\pi[s_k]}(x).$$

Combining, we have proved the claim.

Using the result of the claim, we then have

$$\begin{aligned}\text{peb}_{\pi'}(x) &= \sum_k \text{peb}_{\pi'[s_k]}(x)\\ &\leqq \sum_k \text{peb}_{\pi[s_k]}(x) = \text{peb}_{\pi}(x).\end{aligned}$$

Finally, by Lemma A.4(c),

$$V = \min\,\{\text{peb}_{\pi'}(v_a),\ \text{peb}_{\pi'}(y)\}$$

and we need to show that $V \leqq \text{peb}_{\pi}(z)$, for all nodes z after x in the π sequence. If $z = y$, it is obviously true. Otherwise, it can be verified that for all k, $\text{peb}_{\pi'[s_k]}(v_a) \leqq \text{peb}_{\pi[s_k]}(z)$. This implies that $\text{peb}_{\pi'}(v_a) \leqq \text{peb}_{\pi}(z)$, and hence the result. □

A.3. Monotonicity of Combine algorithm (Theorem 6.3). Theorem 6.3 provides the monotone property of the "Combine" algorithm with respect to subtree orderings. In words, better subtree orderings will yield a better overall ordering by Algorithm 6.1. Before the proof, we introduce a lemma.

LEMMA A.6. *Given two cost sequences with*

$$(\tilde{H}_1, \tilde{V}_1, \cdots, \tilde{H}_{\tilde{r}}, \tilde{V}_{\tilde{r}}) \prec (H_1, V_1, \cdots, H_r, V_r).$$

For $1 \leqq i \leqq \tilde{r}$, *define the function* $f(*)$ *by*

$$f(i) = \min\,\{k \mid \tilde{H}_i \leqq H_k, \tilde{V}_i \leqq V_k\}.$$

If $i < j$, *then* $f(i) \leqq f(j)$.

Proof. By definition of $f(i)$ and Lemma 5.1, we have

$$\tilde{H}_j \leqq H_{f(j)}, \qquad \tilde{V}_i < \tilde{V}_j \leqq V_{f(j)}.$$

If $\tilde{H}_i \leqq H_{f(j)}$, then by definition of $f(i)$, we must have $f(i) \leqq f(j)$. On the other hand, if $H_{f(j)} < \tilde{H}_i$, which together with $\tilde{H}_i \leqq H_{f(i)}$, we have $H_{f(j)} < H_{f(i)}$. By Lemma 5.1, we must have $f(i) < f(j)$. □

Proof of Theorem 6.3. Let $\bar{\psi}_k$ and ψ_k be two subtree orderings on $T[s_k]$, with

$$\text{Pcost}(T[s_k], \bar{\psi}_k) \prec \text{Pcost}(T[s_k], \psi_k).$$

As in the theorem, let

$$\psi = \Phi(\psi_1, \cdots, \psi_k, \cdots, \psi_t).$$

We shall prove the result by first improving on the ordering ψ. We are going to replace subtree segments from $\text{Pcost}(T[s_k], \psi_k)$ in the ordering ψ by those from $\text{Pcost}(T[s_k], \bar{\psi}_k)$. Since $\text{Pcost}(T[s_k], \bar{\psi}_k) \prec \text{Pcost}(T[s_k], \psi_k)$, we can associate each segment from $\bar{\psi}_k$ to one in ψ_k using the mapping $f(*)$ of Lemma A.6. From ψ, construct the new ordering ψ' as follows:

(a) For the ith subtree segment from $\text{Pcost}(T[s_k], \bar{\psi}_k)$, insert it before the corresponding $f(i)$th segment of $\text{Pcost}(T[s_k], \psi_k)$ in ψ;

(b) Remove all subtree segments of $\text{Pcost}(T[s_k], \psi_k)$ from the ordering.

It is clear that ψ' is a segment ordering using the new $\bar{\psi}_k$. Furthermore, by Lemma A.6, it is compatible with the new subtree ordering $\bar{\psi}_k$. We claim that

$$\text{Pcost}(T, \psi') \prec \text{Pcost}(T, \psi).$$

The proof uses the same technique (Lemma A.1) as before and will be skipped.

Finally, let

$$\bar{\pi} = \Phi(\psi_1, \cdots, \bar{\psi}_k, \cdots, \psi_t).$$

By Theorem 6.1, we have

$$\mathrm{Pcost}(T, \bar{\pi}) \prec \mathrm{Pcost}(T, \psi'),$$

and hence the result. □

Acknowledgment. The author would like to thank Professor Andranik Mirzaian for many helpful discussions.

REFERENCES

[1] A. V. Aho, J. E. Hopcroft and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, 1983.

[2] S. A. Cook and R. Sethi, *Storage requirements for deterministic polynomial finite recognizable languages*, J. Comput. System Sci., 13 (1976), pp. 25–37.

[3] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear systems*, ACM Trans. Math Software, 9 (1983), pp. 302–325.

[4] M. R. Garey, R. L. Graham, D. S. Johnson and D. E. Knuth, *Complexity results for bandwidth minimization*, SIAM J. Appl. Math, 34 (1978), pp. 477–495.

[5] J. A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[6] J. Gilbert, T. Lengauer and R. Tarjan, *The pebbling problem is complete in polynomial space*, SIAM J. Comput., 9 (1980), pp. 513–525.

[7] T. Lengauer, *Black-white pebbles and graph separation*, Acta Informatica, 16 (1981), pp. 465–475.

[8] T. Lengauer and R. E. Tarjan, *The space complexity of pebble games on trees*, Inform. Process. Lett., 10 (1980), pp. 184–188.

[9] ———, *Asymptotically tight bounds on time-space trade-offs in a pebble game*, J. ACM, 29 (1982), pp. 1087–1130.

[10] R. J. Lipton and R. E. Tarjan, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 37 (1979), pp. 177–189.

[11] J. W. H. Liu, *On the storage requirement in the out-of-core multi-frontal method for sparse factorization*, Tech. report CS-85-02, Dept. of Computer Science, York University, 1985; ACM Trans. on Math Software, to appear.

[12] ———, *An adaptive general sparse out-of-core scheme for sparse Cholesky factorization*, SIAM Sci. Statist. Comput., 8 (1987), pp. 585–599.

[13] ———, *A compact row storage scheme for sparse Cholesky factors using elimination trees*, ACM Trans. Math Software, 12 (1986), pp. 127–148.

[14] M. C. Loui, *The space complexity of two pebble games on trees*, M.I.T. Technical Report MIT/LCS/TM-133, 1979.

[15] F. Meyer auf der Heide, *A comparison of two variations of a pebble game on graphs*, Theoret. Comput. Sci., 13 (1981), pp. 315–322.

[16] N. Pippenger, *Pebbling*, IBM Research Report RC8258, 1980.

[17] J. K. Reid, TREESOLVE, *a Fortran package for solving large sets of linear finite element equations*, CSS 155, Comput. Sci. Syst. Division, AERE Harwell, Oxfordshire, 1984.

[18] R. Schreiber, *A new implementation of sparse Gaussian elimination*, ACM Trans. Math Software, 8 (1982), pp. 256–276.

[19] R. E. Tarjan, *Data Structures and Network Algorithms*, CBMS-NSF Regional Conference Series in Applied Math, Society for Industrial and Applied Mathematics Publication, 1983.

[20] M. Yannakakis, *A polynomial algorithm for the min cut linear arrangement of trees*, J. ACM, 32 (1985), pp. 950–988.

AN ALGORITHM TO IMPROVE NEARLY ORTHONORMAL SETS OF VECTORS ON A VECTOR PROCESSOR*

BERNARD PHILIPPE†

Abstract. The symmetric orthogonalization, which is obtained from the polar decomposition of a matrix, is optimal. We propose an iterative algorithm to compute this orthogonalization on vector computers. It is especially efficient when the original matrix is near an orthonormal matrix.

Key words. polar decomposition, iterative method, square root, vector computer

AMS(MOS) subject classification. 65F25

Introduction. In the computation of the eigenvectors of a Hermitian matrix, it is necessary to check the orthonormality of the computed vectors, since for close eigenvalues there is an accompanying loss of orthogonality. Usually, especially when the vectors have been computed by inverse iteration, the Gram–Schmidt orthonormalization is performed on the groups of eigenvectors corresponding to close eigenvalues. If the residual is checked before and after this orthogonalization, a loss of accuracy appears. This should not be surprising since Gram–Schmidt orthogonalization corresponds to a QR factorization which depends on the ordering of the vectors. So, instead of a QR factorization, a polar decomposition seems to be preferred because it leads to an orthonormalization which is the best in some sense. This process has been called "Symmetric Orthogonalization" by Lowdin in [LO70].

In this paper, the optimal properties of symmetric orthogonalization are described in § 1. In this section it is also shown that, to orthonormalize a matrix A, it is sufficient to compute $A\,(A^*A)^{-1/2}$.

In § 2, an iterative scheme, which computes $S^{-1/2}$, where S is a Hermitian positive definite matrix, is analyzed and shown to be efficient on vector processors.

In § 3, the complete algorithm for the symmetric orthogonalization is given and experiments are presented.

1. Polar decomposition. In this section, the polar decomposition of a matrix and its application are described. This decomposition is a well-known factorization and a satisfactory presentation is given by Higham in [HA84].

THEOREM 1.1. *Let $A \in \mathbf{C}^{n\times p}$, $n \geqq p$. Then there exists a matrix $U \in \mathbf{C}^{n\times p}$ and a unique Hermitian positive semidefinite matrix $H \in \mathbf{C}^{p\times p}$ such that*

$$A = UH, \qquad U^*U = I_p.$$

If rank (A) = *p then H is positive definite and U is uniquely determined.*

Proof. See [G59]. □

This factorization can be obtained directly from the singular value decomposition of the initial matrix. The SVD insures the existence of unitary matrices $P \in \mathbf{C}^{n\times n}$ and $Q \in \mathbf{C}^{p\times p}$ such that

$$(1.1)^1 \qquad P^*AQ = \underset{\sim}{D}$$

* Received by the editors March 18, 1985; accepted for publication October 16, 1986.

† Institut de Recherche en Informatique et Systèmes Aléatoires, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France.

[1] Notation: $\underset{\sim}{M}$ is the $n \times p$ matrix, obtained by the suppression of the $n - p$ last columns of the matrix $M \in \mathbf{C}^{n\times n}$.

with $D = \operatorname{diag}(\sigma_1, \cdots, \sigma_p, 0, \cdots, 0) \in \mathbf{C}^{n \times n}$ where $0 \leqq \sigma_1 \leqq \cdots \leqq \sigma_p$. Because $A = \underline{P}DQ^*$, then

$$U = \underline{P}Q^* \quad \text{and} \quad H = QD'Q^* = (A^*A)^{1/2},$$

with $D' \in \mathbf{C}^{p \times p}$ and $D' = \operatorname{diag}(\sigma_1, \cdots, \sigma_p)$. When the matrix A is of full rank, the factorization can be performed by using the following theorem.

THEOREM 1.2. *Let $A \in \mathbf{C}^{n \times p}$ with* rank $(A) = p \leqq n$. *Then the polar decomposition $A = UH$ is given by*

$$U = A(A^*A)^{-1/2} \quad \textit{and} \quad H = (A^*A)^{1/2}.$$

Proof. From the SVD (1.1), we have

$$(A^*A)^{-1/2} = QD'^{-1}Q^*$$

and

$$\begin{aligned} A(A^*A)^{-1/2} &= \underline{P}DQ^*(QD'^{-1}Q^*) \\ &= \underline{P}Q^*. \end{aligned}$$ □

There is an algorithm [HB84] that computes $(A^*A)^{-1/2}$ and $(A^*A)^{1/2}$ simultaneously. Here, because we are only looking for the matrix U in the polar decomposition, we use the formulation of Theorem 1.2. Transforming a matrix A into the matrix U is an orthonormalization procedure which we call symmetric orthogonalization. This transformation is different from the usual one which corresponds to the QR factorization. In the following theorem the optimal properties of this symmetric orthogonalization are described.

THEOREM 1.3. *Let $A \in \mathbf{C}^{n \times p}$ with $p \leqq n$ and let $A = UH$ be a polar decomposition. Then*

$$\|A - U\| = \min_{Q \in \mathbf{U}} \|A - Q\|$$

where $\mathbf{U}$ *is the subset of all orthonormal matrices of* $\mathbf{C}^{n \times n}$. *This result is true for both the Euclidean norm and the Frobenius norm.*

Proof. For $p = n$, this result was proved by Fan and Hoffman in [FH55]. Its extension for $p \leqq n$ is straightforward.

2. Computation of the inverse of a square-root.

2.1. Scalar schemes. When Newton's method is used to find the positive root of the polynomial $f(x) = sx^2 - 1$, where $s > 0$, the iterative scheme obtained is

(I) given x_0, $\quad x_{m+1} = (1/2)(x_m + 1/(sx_m))$,

whereas if Newton's method is applied to the function $f(x) = 1/x^2 - s$ the scheme becomes

(II) given x_0, $\quad x_{m+1} = x_m + x_m(1 - sx_m^2)/2$.

When they are convergent, these schemes are quadratically convergent; let $e_m = x_m - s^{-1/2}$ be the error at step m. For (I) and (II) this quantity satisfies the following:

$$e_{m+1} = K_i e_m^2, \qquad i = \mathrm{I}, \mathrm{II}$$

with $K_{\mathrm{I}} = 1/(2x_m)$ and $K_{\mathrm{II}} = -s^{1/2}(s^{1/2}x_m + 2)/2$. Hence, close to the solution, the ratio of the convergence rates of the two schemes is equal to

$$K_{\mathrm{II}}/K_{\mathrm{I}} \simeq -3.$$

The domains of convergence for the two schemes are exhibited in the next result.

PROPOSITION 2.1. *For any positive numbers x_0 and s, the sequence $\{x_m\}$ defined by scheme* (I) *converges quadratically to $s^{-1/2}$.*

For any positive number s, the condition $0 < x_0 < \sqrt{3}s^{-1/2}$ insures that the sequence $\{x_m\}$ defined by scheme (II) *converges quadratically to $s^{-1/2}$.*

Proof. Let us consider the quantity $u_m = s^{1/2}x_m$; the convergence of the sequence $\{x_m\}$ to $s^{-1/2}$ is then equivalent to the convergence of the sequence $\{u_m\}$ to 1. So, the schemes become

(I') given $u_0 = x_0 s^{1/2}$, $u_{m+1} = (u_m + 1/(u_m))/2$

and

(II') given $u_0 = x_0 s^{1/2}$, $u_{m+1} = u_m + u_m(1 - u_m^2)/2$.

The function $u \to g(u) = (u + 1/u)/2$ transforms the interval $(0, +\infty)$ into the interval $[1, +\infty)$ and satisfies the following:

$$u > 1 \text{ implies } 0 < g(u) - 1 = (u-1)^2/(2u) < (u-1)/2.$$

The last inequality proves that scheme (I') is always convergent.

The function $u \to g(u) = u + u(1 - u^2)/2$ transforms the interval $(0, \sqrt{3})$ into the interval $(0, 1]$. If we consider u such that $0 < u < 1$ then

$$0 < 1 - g(u) = (u+2)(1-u)^2/2 < 1 - u.$$

So if $0 < u_0 < \sqrt{3}$ scheme (II') converges. □

In this situation it appears that scheme (I) must be preferred to scheme (II). The generalization of scheme (II) to the matrix situation is much more interesting, since its computation is expressed with matrix multiplications. Moreover, the differences in convergence between (I) and (II) are not as great as in the scalar case.

2.2. Matrix schemes. Let S be a Hermitian positive definite matrix of order p and let $0 < s_1 \leqq \cdots \leqq s_p$ be its eigenvalues. First of all we remark that the only schemes to be considered are those which correspond to the application of the scalar schemes in every eigendirection when the initial guess commutes with S. Because we are only interested in polynomial schemes, we consider the following schemes that are based on (II) of the scalar case:

(Σ_a) given T_0, $T_{m+1} = T_m + \alpha T_m(I - T_m S T_m) + \beta(I - T_m S T_m)T_m$

where α and β are two nonnegative parameters satisfying $\beta = \frac{1}{2} - \alpha$. The quantity $Z_m = I - T_m S T_m$ is called the residual at step m.

THEOREM 2.2. *Let $K(S) > 1$ be the condition number of S* (*ratio of the extremal eigenvalues*). *If $K(S) < 17 + 6\sqrt{8}$ then $S^{-1/2}$ is a point of attraction of the iteration* ($\Sigma_{1/4}$); *this condition becomes $K(S) < 9$ for the iterations* (Σ_0) *or* ($\Sigma_{1/2}$).

Proof. $V = S^{-1/2}$ is a fixed point of the polynomial

$$F_a: T \to T + \alpha T(I - TST) + \beta(I - TST)T.$$

Let us compute its differential application at V. If $T = V + W$ then

$$\begin{aligned} I - TST &= -VSW - WSV + O(W^2) \\ &= -V^{-1}W - WV^{-1} + O(W^2). \end{aligned}$$

Hence

$$T(I - TST) = -W - VWV^{-1} + O(W^2), \qquad (I - TST)T = -V^{-1}WV - W + O(W^2).$$

So, for every matrix W

$$F_a(V+W)-F_a(V)=(1/2)W-\alpha VWV^{-1}-\beta V^{-1}WV+O(W^2).$$

Hence the differential application is given at V by

$$F'_a(V)W=(1/2)W-\alpha VWV^{-1}-\beta V^{-1}WV.$$

To use Ostrowsky's Theorem [OR70], it is necessary to prove that the spectral radius of $F'_a(V)$ is smaller than 1. By using a similarity transformation, we may assume that the matrix V is diagonal:

$$V=\operatorname{diag}(1/\sqrt{s_1},\cdots,1/\sqrt{s_p}).$$

Then it can be proved that the spectrum of $F'_a(V)$ is the set

$$\sigma(F'_a(V))=\{\mu_{ij}\,|\,\mu_{ij}=1/2-\alpha\sqrt{s_i/s_j}-\beta\sqrt{s_j/s_i},\quad i,j=1,n\}.$$

Let λ be any $\sqrt{s_i/s_j}$. We look for the largest interval **I** such that if $\lambda\in\mathbf{I}$ and $1/\lambda\in\mathbf{I}$ then $|1/2-\alpha\lambda-\beta/\lambda|<1$. It is easy to see that this is equivalent to solving the system

$$\alpha\lambda^2-3\lambda/2+\beta<0,\qquad \beta\lambda^2-3\lambda/2+\alpha<0. \tag{2.1}$$

If $\alpha=\beta=1/4$ then (2.1) is equivalent to

$$\lambda^2-6\lambda+1<0$$

and then $\mathbf{I}=(1/\lambda_0,\ \lambda_0)$ with $\lambda_0=3+\sqrt{8}$. Hence $\sigma(F'_{1/4}(V))\subset\mathbf{I}$ is equivalent to $K(S)<(3+\sqrt{8})^2=17+6\sqrt{8}$.

If $\alpha=1/2$ and $\beta=0$ then (2.1) is equivalent to

$$\lambda^2-3\lambda<0,\qquad 1-3\lambda<0$$

and then $\mathbf{I}=(1/3,3)$. Hence $\sigma(F'_{1/2}(V))\subset\mathbf{I}$ is equivalent to $K(S)<9$.

In the same way, the reader can prove that $\sigma(F'_0(V))\subset\mathbf{I}$ is equivalent to $K(S)<9$. □

Remark 2.3. (i) If T_0 is Hermitian, then scheme $(\Sigma_{1/4})$ can be expressed in a better way by

$$\begin{aligned}&\text{given } T_0,\\ &T'_{m+1}=T_m+(1/2)T_m(I-T_mST_m),\\ &T_{m+1}=(1/2)(T'^{*}_{m+1}+T'_{m+1}).\end{aligned}$$

This expression proves that $(\Sigma_{1/4})$ is actually equivalent to using $(\Sigma_{1/2})$ and adding a symmetrization at every step. This formulation is cheaper in terms of operation count than the original one.

(ii) Considering the scheme

$$\text{given } T_0,\qquad T_{m+1}=(1/2)(T_m+(ST_m)^{-1})$$

which is based on the scalar scheme (I), the associated function G defined by

$$G\colon T\to(1/2)(T+(ST)^{-1})$$

has the same differential application as the $(F_0)'$. So, the local convergence of this scheme is only insured if $K(S)<9$. This scheme has been studied by Laasonen in [LA58].

THEOREM 2.4. *Let $\rho(S)$ be the spectral radius of S. If $\mu<(3/\rho(S))^{1/2}$ then the scheme*

$$T_0=\mu I,\qquad T_{m+1}=T_m+(1/2)T_m(I-T_mST_m) \tag{$\bar{\Sigma}$}$$

is quadratically convergent.

Moreover, if $K(S) < 9$ then this scheme is locally stable. This condition can be weakened into $K(S) < 17 + 6\sqrt{8}$ if a symmetrization is performed at every step on T_m.

Proof. By induction, it is clear that every iterate T_m is Hermitian and commutes with S. Because the subspace of the matrices commuting with S is included in the kernel of the differential application which is defined in Theorem 2.2 then scheme $(\bar{\Sigma})$ has a quadratic convergence as soon as it is convergent. In this situation, the scheme is equivalent to using the scalar scheme (II) in every eigendirection. Using the initial guess μ to compute $s_i^{-1/2}$, $i = 1, p$ with the scalar scheme (II) the conditions

$$\mu < \sqrt{3} s_i^{-1/2}, \qquad i = 1, p$$

must be true to insure convergence. These are also sufficient conditions (see Proposition 2.1). So, the first result of the theorem is proved.

If we assume now that this scheme is perturbed by rounding errors, we can no longer insure that T_m commutes with S. The condition $K(S) < 9$ (or $K(S) < 17 + 6\sqrt{8}$ if a symmetrization of T_m occurs at every step) is sufficient to insure that a perturbation due to rounding errors will decrease in the succeeding steps at least in a neighborhood of the solution, since $S^{-1/2}$ is a point of attraction of the iteration (Theorem 2.2). □

PROPOSITION 2.5. *The residual of scheme $(\bar{\Sigma})$, not considering rounding errors, satisfies*

$$Z_{m+1} = (3/4)Z_m^2 + (1/4)Z_m^3. \tag{2.2}$$

Proof.

$$T_{m+1}^2 = (T_m + (1/2)T_m Z_m)^2$$

$$= T_m^2 + (1/4)T_m^2 Z_m^2 + T_m^2 Z_m.$$

Hence

$$Z_{m+1} = I - ST_m^2 - (1/4)ST_m^2 Z_m^2 - ST_m^2 Z_m.$$

When we use $-ST_m^2 = Z_m - I$, the result of the proposition follows. □

Remark 2.6. Formula 2.2 appears to be of interest because it can split the computation of T_{m+1} into two tasks since Z_{m+1} can be evaluated from Z_m only. However, this formula cannot be used repeatedly without updating the residual from its definition $Z_m = I - T_m S T_m$.

3. Computation of the symmetric orthogonalization.

3.1. Application of scheme $(\bar{\Sigma})$. Let us come back to the matrix $A \in \mathbb{C}^{n \times p}$, assuming rank $(A) = p \leqq n$. To orthogonalize this matrix with a symmetric orthogonalization, it is necessary to compute $S^{-1/2}$ where $S = A^*A$ (§ 1). To insure the stability of $(\bar{\Sigma})$ the condition number of S is assumed to be smaller than $(17 + 16\sqrt{8})$.

In order to define an initial guess, the spectral radius $\rho(S)$ of the matrix S has to be estimated. In fact, the ∞-norm is used instead of this spectral radius. From Theorem 2.4 a number μ is then computed:

$$\mu = \sqrt{3/\|S\|_\infty} \leqq \sqrt{3/\rho(S)}.$$

By choosing $T_0 = \mu I$, we ensure the convergence of scheme $(\bar{\Sigma})$. The first iteration can be skipped since it is easy to compute the following:

$$T_1 = (3/2)\mu I - (1/2)\mu^3 S.$$

However, if the matrix $\Delta = S - I$ is small (i.e., $\rho(\Delta) < 1$), the initial guess can be much closer to the solution by choosing the Taylor approximation of order k of $(I + \Delta)^{-1/2}$

$$T_0 = I + \sum_{i=1}^{k} (-1)^i \binom{-1/2}{i} \Delta^i.$$

After m iterations the magnitude of the error is given by

$$T_m - S^{-1/2} = O(\Delta^{(k+1)2^m}).$$

The computation of T_m involves $(k - 1) + 3m$ matrix multiplications. Then, the best order to choose is always smaller than 5. For a required precision ε, an estimation of the best order of the Taylor approximation is given by the author in [PH85] and depends on the ratio $(\log \varepsilon / \log \|\Delta\|_\infty)$.

This algorithm is related to the algorithm which is described in [BB71]: here the matrix $T = (A^*A)^{-1/2}$ is computed before performing the multiplication $A \times T$, hence the iterative part of the algorithm is in $O(p^3)$ flops while it was in $O(np^2)$ in [BB71]. Moreover, the introduction of a symmetrization on the iterate at every stage improves the stability when needed.

3.2. Algorithm. Summarizing the previous considerations, we have the following algorithm.

```
begin
  $S := A^* \times A$ ;
  $\Delta := I - S$ ;
  $\delta := \|\Delta\|_\infty$ ;
  if ($\delta < \varepsilon$) then
        nothing to do ;
    elseif ($\delta < 1$) then
        $k :=$ Taylor approximation order ;
        $T :=$ Taylor approximation of order $k$ ;
        sym := false ;
    else
        $\mu = \sqrt{3/\delta}$ ;
        $T := (3/2)\mu I - (1/2)\mu^3 S$ ;
        sym := true ;
    endif ;
    iter := 0 ;
        loop :
          $\delta 0 := \delta$ ;
          $Z := I - T \times S \times T$ ;
          $\delta := \|Z\|_\infty$ ;
          if ($\delta < \varepsilon$) then exit of the loop endif ;
          if ($\delta > \delta 0$) then divergence endif ;
          iter := iter + 1 ;
          $T := (1/2)T \times (2I + Z)$ ;
          if (sym) then $T := (1/2)(T^* + T)$ endif ;
        endloop ;
  $A := A \times T$ ;
end.
```

TABLE 1
Symmetric orthogonalization on CRAY 1.

$\delta = \|\bar{Q}^T\bar{Q} - I\|_\infty$ before orthog.	With symmetriz.	Taylor order	# of iteration(s)	Elapsed time (unity: 10^{-1} s)
$0.24\,E-3$	No	3	0	0.35
$0.41\,E-3$	No	3	1	0.45
$0.54\,E-2$	No	2	1	0.42
$0.22\,E-1$	No	3	1	0.45
$0.81\,E-1$	No	4	2	0.58
0.39	No	4	3	0.69
$0.27\,E+1$	Yes		14	1.72
$0.34\,E+1$	Yes		7	1.01

To measure the cost of the computation, it is assumed that $A \in \mathbf{R}^{n\times p}$. Following [GVL83] a flop is defined as the amount of computation involved in a triad: $a := a + b \times c$. Then the cost of:

$S := A^T \times A$ is p^2n flops (or $1/2p^2n$ flops if symmetry of S is taken into account),
one iteration of (Σ) is $\approx 3p^3$ flops,
$A := A \times T$ is p^2n flops.

The cost of the Gram–Schmidt orthogonalization is $p^2n + np$ flops. So if $n \gg p$ the symmetric orthogonalization is about twice as expensive as the Gram–Schmidt process, but it is based only on matrix multiplications. If $n = p$ the symmetric orthogonalization becomes more expensive for the computation of $T^{-1/2}$.

An alternative way to compute the symmetric orthogonalization would be to perform the SVD of A or to diagonalize S. In both cases, the number of flops is larger (see [PH85]). Moreover these algorithms are much more difficult to vectorize.

3.3. Experiments. In this section, the results of experiments on a CRAY 1[2] are discussed. An orthogonal matrix $Q \in \mathbf{R}^{201\times 61}$ was constructed from a unitary vector u by $Q = I - 2uu^T$. This matrix Q was randomly perturbed into a matrix $\bar{Q}$ whose column vectors were still normalized (to be in a situation similar to when finding eigenvectors). Both orthogonalizations (symmetric and Gram–Schmidt) were performed on $\bar{Q}$. For symmetric orthogonalization, the results for different magnitudes of perturbation are exhibited in Table 1. For each run, the algorithm is defined by the value of the quantity δ. If δ is smaller than 1 then the initial guess is obtained by a Taylor expansion whose order is given in Table 1. If δ is larger than 1 then the initial guess is μI, where μ is computed from δ (§ 3.2). In this last case, a symmetrization on the iterate occurs at every stage.

After orthogonalization, the residual $\|\bar{Q}^T\bar{Q} - I\|_\infty$ was always in the range $[10^{-13}, 10^{-12}]$. The elapsed time for the Gram–Schmidt process was 0.59×10^{-1} s.

For each run, the distance between the perturbed matrix and its orthogonalized matrix was very close to the residual given in the first column of Table 1 for Gram–Schmidt. For the symmetric orthogonalization, the distance was only half of this residual. Some cases of divergence were obtained with perturbation of larger magnitude. These cases correspond to matrices $\bar{Q}^T\bar{Q}$ with a small eigenvalue which implies a large condition number. In these situations, the solution was almost reached before the rounding errors became important because of increasing magnitude at every iteration.

[2] This CRAY 1 is managed by the Conseil Scientifique du Centre de Calcul Vectoriel pour la Recherche, Palaiseau, France.

Conclusion. Even when the result of a computation should be an orthonormal set of vectors (e.g., for the eigenvectors of a Hermitian matrix), there is often a loss of orthogonality which occurs due to rounding errors. In this situation the orthogonalization process should preserve the quality of the original set. As has been proved, the symmetric orthogonalization is optimal. The iterative scheme which is proposed in this paper is efficient on vector processors since it uses only matrix multiplications. This scheme is numerically stable when the ratio of the extremal singular values is smaller than $3 + \sqrt{8}$.

Acknowledgment. The author would like to thank the referees for their valuable criticism of a first version of this paper. He is also grateful to B. Parlett and A. Sameh for their helpful remarks.

REFERENCES

[BB71] A. Bjorck and C. Bowie, *An iterative algorithm for computing the best estimate of an orthogonal matrix*, SIAM J. Numer. Anal., 8 (1971), pp. 358–364.

[FH55] K. Fan and A. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[G59] F. R. Gantmacher, *The Theory of Matrices*, Volume One, Chelsea, New York, 1959.

[GVL83] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[HA84] N. J. Higham, *Computing the polar decomposition—with applications*, Numerical Analysis Report No. 94, Univ. of Manchester, Manchester, England, 1984.

[HB84] ———, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–550.

[LA58] P. Laasonen, *On the iterative solution of the matrix equation $AX^2 - I = 0$*, Math. Tables Aids Comput., 12 (1958), pp. 109–116.

[LO70] P. Lowdin, *Advances in Quantum Chemistry*, Vol. 5, Academic Press, New York, 1970.

[OR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[PA80] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ 1980.

[PH85] B. Philippe, *Approximating the square root of the inverse of a matrix*, Cedar document No. 108, CSRD, Univ. of Illinois, Urbana, IL 1985.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

NETWORK RESILIENCE*

CHARLES J. COLBOURN†

Abstract. The resilience of a network is a measure of its reliability; it is the expected number of node pairs which can communicate. The resilience of an n-vertex series-parallel network can be computed in $O(n^2)$ time. The algorithm employs the recursive structure of maximal series-parallel networks. In contrast to this, computing the resilience of a planar network is shown to be #P-complete.

Key words. network reliability, network resilience, #P-completeness, series-parallel network, planar network

AMS(MOS) subject classifications. 68C25, 68E99

1. Preliminaries. The design of computer communication networks has as a primary goal the production of *reliable* networks. The demand for reliable systems has led naturally to the development of many formal definitions of reliability; however, one definition is prevalent [12], [1]. A network is modelled as a probabilistic graph (V, E); V is a set of nodes representing communication sites, and E is a set of undirected edges representing communication links between pairs of nodes. Edges have an associated *success probability*; for our purposes, the probability is a fixed precision real number to avoid difficulties with "infinite precision" arithmetic. With this model, the *all-terminal reliability* is the probability that every pair of nodes can communicate, when edge failures are assumed to be statistically independent. A large body of research has been undertaken on this measure; for general networks, it is #P-complete to compute it [7]. This has led to the development of efficiently computable bounds on the reliability [3], and to efficient solutions in restricted cases, notably series-parallel networks [11], [14]. A related measure, the *two-terminal reliability* is the probability that two specified modes can communicate.

Van Slyke and Frank [12] remark that usage of the all-terminal reliability, although widespread, is not appropriate in certain applications. Often, one is not concerned that the network be connected, but rather that "most" potential communicating vertex pairs remain connected. They suggest that in many applications, a more appropriate measure of reliability is the *expected number of node pairs* which can communicate; we term this statistic the *resilience* of the network, since it captures (in part) the network's capacity to withstand failures. We denote by Res (G) the resilience of a network G.

One of the more difficult areas of investigation is the relation between all-terminal reliability and resilience; this has been hampered by the lack of practical algorithms for computing resilience in any nontrivial class of networks. Hence, although all-terminal reliability in series-parallel networks admits a linear time solution [14], no corresponding result is known for resilience. However, such a result would be useful, for example, in studying the relation between reliability and resilience; some preliminary work in this direction appears in [2]. Moreover, much research in network design and analysis has been devoted to series-parallel networks (see, for example, [4], [9]).

In this paper, we develop an $O(n^2)$ algorithm for the resilience of an n-vertex series-parallel network. In addition to the practical applications, the algorithm is of interest for its use of a class of recursively defined graphs, the 2-trees. A graph is *series-parallel* when it contains no subgraph homeomorphic to K_4; an easy consequence is that all series-

* Received by the editors April 14, 1986; accepted for publication October 15, 1986. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant A0579.

† Computer Communications Networks Group, Department of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

parallel graphs are planar, and all outerplanar graphs are series-parallel. An equivalent definition is the class of graphs each of whose biconnected components is reducible to an edge via series and parallel reductions. Here, a series reduction involves removing a degree two node and adding an edge between its neighbours. A parallel reduction involves replacing parallel or multiple edges with a single edge. A *2-tree* is an n-node graph which is either a single edge (i.e., K_2), or obtained from an $n-1$-node 2-tree by selecting an adjacent pair $\{x, y\}$ of nodes and adding a new node z along with the edges $\{x, z\}$ and $\{y, z\}$. 2-trees are easily seen to be series-parallel; in fact, Wald and Colbourn [13] showed that every n-vertex series-parallel network is a subgraph of an n-vertex 2-tree. Moreover, they described a linear time algorithm for determining a set of edges which complete a series-parallel network to a 2-tree.

In reliability applications, these added edges each have assigned probability zero; in this way, we need only compute resilience for 2-trees. This transformation of problems on series-parallel networks to problems on 2-trees has been successful in a number of problems, notably Steiner tree [13], reliability [14], optimum communication spanning tree [5], and Steiner 2-edge-connected subnetworks [15]. We adopt this approach here, in order to exploit the recursive structure of 2-trees. The existence of a polynomial time algorithm here is contrasted with the situation for planar networks, where we establish that computing resilience is #P-complete.

2. Remarks on resilience. In preface to the algorithm proper, we first examine the structure of Res (G) for a network G. Res (G) is the expected number of node pairs which can communicate; to be precise, for a subgraph $H \subseteq G$, let Prob (H) be the probability that all edges of H succeed and all edges of G-H fail, and let Pairs (H) be the number of communicating node pairs in H. Then

$$\text{Res}\,(G)=\sum_{H\subseteq G}\text{Prob}\,(H)\times\text{Pairs}\,(H).$$

Now define Pairs (H, x, y) to be 1 if x can communicate with y in H, and 0 otherwise. Then

$$\text{Res}\,(G)=\sum_{H\subseteq G}\text{Prob}\,(H)\times\sum_{y>x}\text{Pairs}\,(H,x,y).$$

Simple algebraic manipulation gives

$$\text{Res}\,(G)=\sum_{y>x}\sum_{H\subseteq G}\text{Prob}\,(H)\times\text{Pairs}\,(H,x,y).$$

This can now be viewed as the summation of $\binom{n}{2}$ measures of *two-terminal* reliability, one for each unordered pair of nodes. Hence it is sufficient to solve a two-terminal reliability problem for $O(n^2)$ pairs, summing the results. Wald and Colbourn's algorithm [14] solves two-terminal reliability on series-parallel networks in linear time, which yields an $O(n^3)$ algorithm for resilience.

The improvement of this to $O(n^2)$ time hinges on a closely related observation. If we denote

$$E(x)=\sum_{y\neq x}\sum_{H\subseteq G}\text{Prob}\,(H)\times\text{Pairs}\,(H,x,y)$$

then

$$\text{Res}\,(G)=\frac{1}{2}\sum_{x}E(x).$$

Hence, an $O(n^2)$ algorithm overall results from a linear time algorithm for computing

$E(x)$, the expected number of nodes which can communicate with a specified node x. This is the strategy which we employ.

3. Reducing 2-trees. The recursive structure of 2-trees leads to an efficient reduction procedure, by repeated elimination of degree 2 nodes until the remaining graph is a single edge. This reduction leads to quite a general class of algorithms, as follows. With each edge $\{x, y\}$, we associate a number of statistics (these may be costs, weights, probabilities, and the like). The statistics associated with an edge at a given point in the reduction always refer to the subgraph which has thus far been "reduced onto" this edge. More specifically, the initial graph associated with an edge is simply the edge itself. When a degree 2 node z is eliminated in the reduction, we locate the neighbours x and y of z. Prior to the reduction, three subgraphs are induced:

L, the subgraph reduced onto $\{x, z\}$;
R, the subgraph reduced onto $\{y, z\}$;
M, the subgraph reduced onto $\{x, y\}$.

When z is removed, the subgraph reduced onto $\{x, y\}$ is updated to $L \cup M \cup R$; in the process, the statistics associated with $\{x, y\}$ must be updated using the statistics for L, M, and R. The "trick" here is associating meaningful statistics with each edge as the reduction proceeds, to enable us to recover the result from the statistics remaining when only a single edge is left.

This recursive reduction has been used before (e.g., [13], [14], [15]). For the computation of resilience, we need to make a small modification. It is easily verified that every 2-tree has at least two nodes of degree 2. Hence at every reduction step in the algorithm, we have a choice of two or more degree 2 nodes to remove. We can therefore ensure that a specified node s is never deleted, and hence remains in the final edge (in fact, one can reduce a 2-tree onto *any* one of its edges, but we do not need this powerful a result here).

With this general framework in mind, we introduce the necessary statistics to compute $E(s)$, the expected number of node pairs involving node s. Each statistic (as above) is defined with respect to a subgraph reduced onto an edge; such a subgraph H has two identified vertices a and b where $\{a, b\}$ is the edge onto which H has been reduced. Then we define

(1) $\Pr_H[a \sim b]$ is the probability that a can communicate with b in H.
(2) $E_H(a)$ is the expected number of nodes which can communicate with a but *not* with b in H. This expected number *includes* a when a and b cannot communicate.
(3) $E_H(b)$ is the expected number of nodes which can communicate with b but not with a in H. This expected number includes b when a and b cannot communicate.
(4) $E_H(ab)$ is the expected number of nodes which can communicate with both a and b. This expected number includes both a and b when they can communicate.

Given these measures, there are three issues which must be addressed: initialization, reduction, and termination. Initially, we must define the measures simply for edges.

LEMMA 3.1 (initialization). *Suppose $H = e = \{x, y\}$, with success probability p_e. Then*

(1) $\Pr_H[x \sim y] = p_e$;
(2) $E_H(x) = 1 - p_e$;
(3) $E_H(y) = 1 - p_e$;
(4) $E_H(xy) = 2 \times p_e$.

Proof. Rule (1) is trivial. Rules (2) and (3) follow from the observation that exactly one node can communicate with $x(y)$ when edge e is down, namely $x(y)$ itself. When e is up, both nodes can communicate with *both* x and y. Similarly, (4) follows from the observation that two nodes can communicate with x and y when e is up. □

The reduction step is naturally more complicated.

LEMMA 3.2 (reduction). *Let G be a (partially reduced) 2-tree; let z be a degree 2 node in G with neighbours x and y. Let L, R and M be the subgraphs reduced so far onto $\{x, z\}$, $\{y, z\}$, and $\{x, y\}$, respectively. Let $B = L \cup R$, with identified nodes x and y; then let $C = B \cup M$. Then*

(1) $\mathrm{Pr}_B[x \sim y] = \mathrm{Pr}_L[x \sim z] \times \mathrm{Pr}_R[z \sim y]$;

(2) $E_B(x) = E_L(x) + \mathrm{Pr}_L[x \sim z] \times E_R(z) + \mathrm{Pr}_R[z \not\sim y] \times E_L(xz) - \mathrm{Pr}_L[x \sim z] \times \mathrm{Pr}_R[z \not\sim y]$;

(3) $E_B(y)$ *is symmetric to* $E_B(x)$;

(4) $E_B(xy) = \mathrm{Pr}_L[x \sim z] \times E_R(zy) + \mathrm{Pr}_R[z \sim y] \times E_L(xz) - \mathrm{Pr}_L[x \sim z] \times \mathrm{Pr}_R[z \sim y]$;

(1′) $\mathrm{Pr}_C[x \sim y] = \mathrm{Pr}_B[x \sim y] + \mathrm{Pr}_M[x \sim y] - \mathrm{Pr}_B[x \sim y] \times \mathrm{Pr}_M[x \sim y]$;

(2′) $E_C(x) = \mathrm{Pr}_M[x \not\sim y] \times E_B(x) + \mathrm{Pr}_B[x \not\sim y] \times E_M(x) - \mathrm{Pr}_B[x \not\sim y] \times \mathrm{Pr}_M[x \not\sim y]$;

(3′) $E_C(y)$ *is symmetric to* (2′);

(4′) $E_C(xy) = E_B(xy) + \mathrm{Pr}_M[x \sim y] \times (E_B(x) + E_B(y)) + E_M(xy) + \mathrm{Pr}_B[x \sim y] \times (E_M(x) + E_M(y)) - 2 \times [1 - \mathrm{Pr}_B[x \not\sim y] \times \mathrm{Pr}_M[x \not\sim y]]$.

Proof. Rules (1) and (1′) follow from the observation that L and R, and B and M are edge-disjoint, and failures are statistically independent. The remainder are verified by exhaustive analysis of cases. This verification is assisted by observing that

$$E_G(x) = \sum_{s \in G} \Pr[x \sim s \text{ and } s \not\sim y] \quad \text{and} \quad E_G(xy) \equiv \sum_{s \in G} \Pr[x \sim s \text{ and } s \sim y].$$

Then to verify (2) (and thus also (3)), consider the contribution of each node in B. A node in L contributes when one of the following holds: x can't reach z in L (independent of the state of R), or x can reach z in L but z cannot reach y in R. This contribution summed over all nodes in L gives the first and third terms in (2). A node in R contributes only if x communicates with z in L and z cannot reach y in R. Summed over all nodes in R, this gives the second term. The final (negative) term arises because z is in both L and R, and hence is accounted for twice. Hence the probability for node z must be subtracted out once.

Each rule is verified in this manner, by determining the contribution of each node and summing appropriately. In each case, a negative term arises due to overcounting for the nodes which appear in both L and R or both B and M. Only in rule (4′) does the correction involve two nodes rather than one. □

The rules in Lemma 3.2 can be applied to reduce any 2-tree to a single edge. At this point, the desired result must be recovered.

LEMMA 3.3 (termination). *Let G be a 2-tree consisting of a single edge $\{x, y\}$ with associated measures computed using Lemmata 3.1 and 3.2. Then*

$$E(x) = E_G(x) + E_G(xy) - 1.$$

Proof. $E_G(x) + E_G(xy) = \sum_{s \in G} \Pr[x \sim s]$. Now $\Pr[x \sim x] = 1$, and hence

$$E(x) = \sum_{\substack{x \neq s \\ x \in G}} \Pr[x \sim s] = E_G(x) + E_G(xy) - 1. \qquad \square$$

Employing the reductions developed in Lemmata 3.1–3.3 along with the general reduction technique developed earlier establishes the following:

THEOREM 3.4. *The expected number of nodes $E(x)$ able to communicate with a specified node x in a series-parallel network can be computed in linear time.*

Proof. Correctness follows from Lemmata 3.1–3.3. Timing can be verified as follows. Completion of the series-parallel network to a 2-tree requires linear time [13]. Subsequent identification of degree 2 nodes, and initialization of the measure on each edge requires linear time (recall that a 2-tree has exactly $2n-3$ edges [13]). A linear number of reductions are required, and each takes constant time. Selecting the degree 2 node to remove is done by maintaining a stack of all degree 2 nodes. Having located its neighbours, all of the reductions can easily be performed in constant time. A potential difficulty arises in finding the neighbours of the degree 2 node, since its entry in the adjacency list may still contain nodes which were previously deleted. Nevertheless, each edge is scanned at most twice in this way, so the overall contribution is linear. Finally, termination requires only constant time. □

COROLLARY 3.5. *The resilience* Res (G) *for a series-parallel network G on n nodes can be computed in $O(n^2)$ time.* □

4. Resilience of planar networks is #P-complete. One might hope to generalize the results for series-parallel networks to obtain polynomial time algorithms for larger classes. Not surprisingly, using techniques outlined in [6], a polynomial time algorithm can be derived for partial k-trees with any fixed k. While of interest, this does not suggest any method for handling graphs arising in networks applications, such as planar graphs. We establish here that the possibility of extending the ideas here to planar graphs are very remote, by establishing that computing resilience for planar networks is #P-complete. We rely on a theorem due to Provan [8]:

THEOREM 4.1. *Computing two-terminal reliability for planar networks is* #P-*complete, even when all edge operation probabilities have the same value, p.*

Our strategy is to reduce two-terminal reliability for planar networks to resilience for planar networks. Let $G=(V,E)$ be a planar graph, and let s and t be the two nodes required to communicate. We define a family of graphs $G_{i,j}$ as follows. $G_{i,j}$ has vertex set $V \cup \{x_1, \cdots, x_i\} \cup \{y_1, \cdots, y_j\}$, and edge set

$$E \cup \{\{x_k, s\} \mid 1 \leqq k \leqq i\} \cup \{\{y_k, t\} \mid 1 \leqq k \leqq j\}.$$

We assume that each edge has the same probability p. We denote by $\mathrm{Rel}_2\,(G;\, s,\, t)$ the two-terminal reliability of s and t in G.

LEMMA 4.2.

$$\mathrm{Res}\,(G_{1,0}) = \mathrm{Res}\,(G) + p(1+E(s)),$$

$$\mathrm{Res}\,(G_{0,1}) = \mathrm{Res}\,(G) + p(1+E(t)),$$

$$\mathrm{Res}\,(G_{1,1}) = \mathrm{Res}\,(G) + p(1+E(s)) + p(1+E(t)) + p^2\mathrm{Rel}_2\,(G; s, t).$$

Proof. Res $(G_{i,j})$ is the sum of the two-terminal reliabilities of each unordered pair of nodes in $G_{i,j}$. The primary observation employed is that in the computation of the two-terminal reliability from a to b, edges not appearing on an a, b-path are *irrelevant*, and can be deleted without affecting the reliability (see, for example, [10]). A second (trivial but necessary) observation is that if G has a node x which appears on every a, b-path, $\mathrm{Rel}_2\,(G;\, a,\, b) = \mathrm{Rel}_2\,(G;\, a,\, x) \times \mathrm{Rel}_2\,(G;\, x,\, b)$. □

This simple lemma enables us to prove:

THEOREM 4.3. *Resilience of planar networks is* #P-*complete.*

Proof. Membership in #P is straightforward. To show completeness, we reduce two-terminal reliability of planar networks to resilience. For an instance G, p of the two-terminal reliability problem, form $G_{1,0}$, $G_{0,1}$ and $G_{1,1}$ as above. Using Lemma 4.2, observe that

$$\mathrm{Res}\,(G_{0,1}) = \mathrm{Res}\,(G) + p \sum_{v \in V} \mathrm{Rel}_2\,(G; v, t),$$

$$\mathrm{Res}\,(G_{1,0}) = \mathrm{Res}\,(G) + p \sum_{v \in V} \mathrm{Rel}_2\,(G; v, s),$$

$$\mathrm{Res}\,(G_{1,1}) = \mathrm{Res}\,(G) + p \sum_{v \in V} \mathrm{Rel}_2\,(G; v, t) + p \sum_{v \in V} \mathrm{Rel}_2\,(G; v, s) + p^2 \mathrm{Rel}_2\,(G; s, t).$$

Compute Res (G) itself as well, and subtract it from each of the given expressions. The equations for $G_{1,0}$ and $G_{0,1}$ then determine the third term in the expression for $G_{1,1}$, namely $p^2\mathrm{Rel}_2$ $(G;\ s,\ t)$. But then computing the resilience of four networks whose size is polynomial in the size of G enables us to compute the two-terminal reliability of G, and the proof is complete. □

The result proved here is what one would expect, and suggests once again the importance of relating resilience and all-terminal reliability, since all-terminal reliability for planar networks remains open.

5. Conclusions. Series-parallel networks form an important, but small, practical class of networks; we have shown here that an embedding into the recursive graph family of 2-trees enables us to solve a difficult reliability problem, computing the resilience. The algorithm is straightforward to implement, and employs a very useful general strategy for solving problems on series-parallel networks. The extension to other classes of practical networks seems unlikely to succeed, in view of the #P-completeness of resilience, even for planar networks.

Of most interest here is the relation between all-terminal reliability and resilience. One expects resilience to be a finer measure of the usefulness of a network. However, much future research will be required to determine what relationships (if any) obtain between the two measures.

Acknowledgments. Thanks to David Johnson and to one of the referees for helpful comments.

REFERENCES

[1] M. O. Ball and J. S. Provan, *Calculating bounds on reachability and connectedness in stochastic networks*, Networks, 13 (1983), pp. 253–278.

[2] B. N. Clark, E. M. Neufeld and C. J. Colbourn, *Maximizing communicating vertex pairs in series-parallel networks*, IEEE Trans. Reliability, R-35 (1986), pp. 247–251.

[3] C. J. Colbourn and D. D. Harms, *Bounding all-terminal reliability in computer networks*, Technical Report E-123, Computer Communications Networks Group, University of Waterloo, 1985.

[4] R. J. Duffin, *Topology of series-parallel networks*, J. Math. Appl., 10 (1965), pp. 303–318.

[5] E. S. El-Mallah and C. J. Colbourn, *Optimum communication spanning trees in series-parallel graphs*, SIAM J. Comput., 14 (1985), pp. 915–925.

[6] ———, *Partial k-tree algorithms*, Utilitas Math., to appear.

[7] J. S. Provan and M. O. Ball, *The complexity of counting cuts and of computing the probability that a graph is connected*, SIAM J. Comput., 12 (1983), pp. 777–778.

[8] J. S. Provan, *The complexity of reliability computations in planar and acyclic networks*, SIAM J. Comput., 15 (1986), pp. 694–702.

[9] A. Rosenthal, *Series-parallel reduction for difficult measures of network reliability*, Networks, 11 (1981), pp. 323–334.

[10] A. Satyanarayana and A. Prabhakar, *New topological formula and rapid algorithm for reliability analysis of complex networks*, IEEE Trans. Reliability, R-27 (1978), pp. 82–100.

[11] A. Satyanarayana and R. K. Wood, *A linear time algorithm for computing k-terminal reliability in series-parallel networks*, SIAM J. Comput., 14 (1985), pp. 818–832.

[12] R. M. Van Slyke and H. Frank, *Network reliability analysis* I, Networks, 1 (1972), pp. 279–290.

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

[14] ———, *Steiner trees in probabilistic networks*, Microelectronics and Reliability, 23 (1983), pp. 837–840.

[15] P. Winter, *Generalized Steiner problem in series-parallel networks*, to appear.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

009

QUASI-MONOTONIC SEQUENCES: THEORY, ALGORITHMS AND APPLICATIONS*

ANDREZJ EHRENFEUCHT†, JEFFREY HAEMER‡ AND DAVID HAUSSLER§

Abstract. We present a simple algebraic theory which allows us to solve a variety of combinatorial problems, including the problem of finding convex hulls in two dimensions, the "Trip Around the Moon" problem, a version of the ballot problem, and the problem of enumerating and randomly generating ordered trees of a given size. Individual problems are solved by applying general algorithms and theorems developed within this algebraic theory.

Key words. amortized complexity, trees, convex hulls, ballot problems, cyclic conjugates, Lyndon words, factorizations

AMS(MOS) subject classifications. 06F99, 68R05, 05C05

Introduction. Imagine yourself standing between a pair of adjacent elements in a sequence of reals. If the sequence is monotonically increasing, then regardless of your exact position within the sequence, every element to your left is less than every element to your right (or vice versa). In this paper we develop the theory of a different, but related kind of sequence. These are *quasi-increasing* sequences, in which the "average" of all the elements to your left is less than the "average" of the elements to your right (or vice versa). Consider as an example an "ideal business year," in which the average of the monthly profits for the remaining months of the year always exceeds the current average.

We develop the theory of quasi-increasing sequences (and the other forms of quasi-monotonic sequences) using the general notion of an *averaging function.* An averaging function is a mapping μ from nonempty sequences over an arbitrary set into some linearly ordered range which satisfies one basic axiom: for any two sequences U and V, $\mu(UV)$ and $\mu(VU)$ must lie between $\mu(U)$ and $\mu(V)$. (Note that we do not demand that $\mu(UV) = \mu(VU)$.) Many of the commonly used measures of central tendency satisfy this basic condition, and are thus averaging functions in the sense that we use this term.

With respect to a given averaging function μ, a sequence S is quasi-increasing if $\mu(U) < \mu(V)$ for every pair of nonempty sequences U and V such that $S = UV$. Quasi-nondecreasing, quasi-decreasing and quasi-nonincreasing sequences are defined analogously. Section 1 gives a brief introduction to the theory of quasi-monotonic sequences. This theory is related to the theory of Viennot factorizations (see [Lot83], [Vie78]), but our basic approach is somewhat different. As in [Vie78], we obtain interesting generalizations of earlier work in [Spi56], and of the work on Lyndon words (see [Lot83]). However, our primary concentration is on the applications of the theory to the solution of various combinatorial problems. These include the problem of finding convex hulls in the plane (see e.g. [Pre79]), the "Trip Around the Moon" problem ([Gra83]), a generalized version of the ballot problem [Tak67], and the problem of enumerating and randomly generating ordered trees of a given size (see e.g. [Der80]). Detailed descriptions of these problems can be found in §§ 2.3, 3.2, 3.4 and 3.5, respectively. While these problems are certainly not new or unsolved, until this point they have not been cast and

* Received by the editors May 31, 1984; accepted for publication (in revised form) October 30, 1986.

† Department of Computer Science, University of Colorado, Boulder, Colorado 80302. The work of this author was supported in part by National Science Foundation grant MCS-83-05245.

‡ Interactive Systems, Inc., Boulder, Colorado 80302.

§ Department of Computer and Information Science, University of California, Santa Cruz, California 95064. The work of this author was supported in part by National Science Foundation grant IST-83-17918.

solved within a general algebraic framework. Sections 2 and 3 are devoted to this task. Each begins with a general result in the theory of quasi-monotonic sequences, followed by a development of an algorithm based on this result with explicit applications.

In § 2 we demonstrate that for any fixed averaging function μ, every sequence can be uniquely decomposed into a series of maximal quasi-increasing segments, called *upward trends* (Theorem 2.1.1). We give a general algorithm for computing this decomposition, prove that it is correct and demonstrate that it is optimal (linear time) for a certain class of averaging functions which we call *constant time merging.* Using a suitable constant time merging averaging function, the convex hull of a sorted sequence of points in the plane can be viewed as a pair of decompositions of the sequence of line segments between adjacent points, where one decomposition gives the upward trends and the other gives the downward trends. Since we have a linear time decomposition algorithm for this averaging function, this gives a linear time algorithm for finding the convex hull of a set of points sorted on one coordinate, and an $O(n \log n)$ algorithm if initial sorting is required. Both algorithms are optimal [Yao79].

In § 3 we demonstrate that every sequence has a cyclic conjugate that is quasi-nondecreasing (Theorem 3.1.2). We give an algorithm for finding this cyclic conjugate which is also optimal for constant time merging averaging functions. This algorithm can be used to solve the "Trip Around the Moon" problem, variants of which are discussed in [Tak67] and [Dvo80] in the context of queuing theory and data storage and retrieval techniques for magnetic bubble memories.

Using a canonical mapping from ordered trees to sequences given in [Read72], an extension of the above result (Theorem 3.3.4) can also be used to obtain formulas enumerating the number of various types of ordered trees by size [Der80], and the above algorithm can be used to randomly generate ordered trees of various sizes and types. These results are given in § 3.4. In addition to these applications, we can also obtain solutions to some generalized forms of the classic ballot problem (see § 3.5).

1. Basics.

1.1. Notation. Throughout this paper, italicized upper-case letters denote finite sequences, and the corresponding lower-case letters denote their elements. Thus a typical sequence is denoted $S = s_1 \cdots s_n$. The *length* of S is n, denoted $|S|$. Sets will be denoted with upper case Greek letters. If Ω is a finite set, then $|\Omega|$ denotes the cardinality of Ω. For any set Σ, Σ^+ denotes the set of all nonempty sequences formed from the elements of Σ. If U and V are sequences, UV denotes the sequence resulting from the concatenation of U and V, and U^k denotes the sequence resulting from the concatenation of U with itself k times. If $S = UWV$, then W is a *segment* of S. If, in addition, U is the empty sequence, then W is a *prefix* of S and if V is the empty sequence, then W is a *suffix* of S. Any segment of S is *proper* if it is not empty and it is not all of S.

1.2. Averaging systems. The basic framework underlying the theory of quasi-monotonic sequences can be described as follows.

DEFINITION. Let Σ and Γ be arbitrary sets, the latter being linearly ordered by a relation $\leqq$. Let μ be an arbitrary function from Σ^+ into Γ which satisfies the following, where U and V are arbitrary sequences in Σ^+.

Interpolation property.

If $\mu(U) < \mu(V)$ then $\mu(U) < \mu(UV), \mu(VU) < \mu(V)$, and

if $\mu(U) = \mu(V)$ then $\mu(U) = \mu(UV) = \mu(VU) = \mu(V)$.

Such a function μ is called an *averaging function* and the system Σ, Γ, $\leqq$, μ is called an *averaging system.*

This general notion of an averaging function encompasses several measures of central tendency which are commonly used, and some not so commonly used. The following are a few examples of averaging functions.

DEFINITION. Given a sequence of real numbers $S = s_1 \cdots s_n$, the *arithmetic mean* of S is

$$\frac{\sum_{i=1}^{n} s_i}{n}.$$

The *geometric mean* of S is $(s_1 s_2 \cdots s_n)^{1/n}$. The *harmonic mean* of S is

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{s_i}}.$$

We restrict the geometric and harmonic means to sequences of positive reals. For any real number $\alpha > 0$, the *α-weighted mean* of S is

$$(s_1 + \alpha s_2 + \alpha^2 s_3 + \cdots + \alpha^{n-1} s_n)/(1 + \alpha + \alpha^2 + \cdots + \alpha^{n-1}).$$

For any sequence of pairs of real numbers

$$T = (x_1, y_1) \cdots (x_n, y_n) \quad \text{where } x_i > 0, 1 \leqq i \leqq n,$$

the *gradient mean* of T is $\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$.

Note that the arithmetic mean is a special case of the α-weighted mean with $\alpha = 1$, and (essentially) a special case of the gradient mean with $x_i = 1$ for all i, $1 \leqq i \leqq n$. Another useful special case of the α-weighted mean is obtained by taking α to be infinitesimal. For μ defined in this manner, given nonempty sequences of reals U and V we have $\mu(U) < \mu(V)$ if and only if UV lexicographically precedes VU, i.e., if and only if $UV = XaZ$ and $VU = XbZ'$ where a and b are reals with $a < b$ and X, Z, Z' are (possibly empty) sequences of reals. We will call this the *lexicographic mean*. This is a good example of a mean which depends on the order of the elements in the sequence, in contrast to the other functions given above.

The fact that all of the above functions are averaging functions rests primarily on one elementary arithmetic result.

LEMMA 1.2.1. *For any real numbers a, b, c, d with $b, d > 0$,*

(1) *if $a/b < c/d$ then $a/b < (a + c)/(b + d) < c/d$, and*

(2) *if $a/b = c/d$ then $a/b = (a + c)/(b + d) = c/d$.*

Proof. This follows easily from well-known arithmetic rules for manipulating fractions. □

LEMMA 1.2.2. *The Interpolation Property holds for all of the above means.*

Proof. That the arithmetic, harmonic and gradient means satisfy the Interpolation Property (under the restrictions given in their definitions) follows directly from the above lemma. In the case of the α-weighted mean, we notice that if $X = x_1 \cdots x_n$ and $Y = y_1 \cdots y_m$, then $\mu(XY)$ is

$$\frac{(x_1 + \cdots + \alpha^{n-1} x_n) + \alpha^n (y_1 + \cdots + \alpha^{m-1} y_m)}{(1 + \cdots + \alpha^{n-1}) + \alpha^n (1 + \cdots + \alpha^{m-1})}.$$

Hence again we can use the above lemma.

That the geometric mean satisfies the Interpolation Property actually follows from the fact that the arithmetic mean satisfies this property. This is because the logarithm of the geometric mean of a sequence of positive reals is the arithmetic mean of their logarithms, and the logarithm is monotonic. □

Since it is our intention to proceed as rapidly as possible to the theory of quasi-monotonic sequences and its applications, we will not give an extensive axiomatic treatment of the theory of averaging systems here. However, we will pause to note a few general properties of averaging systems which will be useful in what follows. From this point on, μ will denote an arbitrary averaging function, and all sequences will be assumed to be sequences over the domain of μ (i.e. nonempty sequences) unless otherwise noted.

One property of averaging systems that agrees well with our intuitive notion of an "average" is the following.

LEMMA 1.2.3 (Balancing Lemma). *Let $U_1, U_2, \cdots, U_k, V_1, V_2, \cdots, V_l$ be sequences, where $k, l > 0$. If $\mu(U_i) \leqq \mu(V_j)$ for all i and j, $1 \leqq i \leqq k$ and $1 \leqq j \leqq l$, then $\mu(U_1 \cdots U_k) \leqq \mu(V_1 \cdots V_l)$. If in addition $\mu(U_i) < \mu(V_j)$ for some i and j, $1 \leqq i \leqq k$ and $1 \leqq j \leqq l$, then $\mu(U_1 \cdots U_k) < \mu(V_1 \cdots V_l)$.*

Proof. Let $\alpha = \max_{1 \leqq i \leqq k} \mu(U_i)$ and let $\beta = \min_{1 \leqq j \leqq l} \mu(V_j)$. By the Interpolation Property (repeatedly), it follows that $\mu(U_1 \cdots U_k) \leqq \alpha$ and $\mu(V_1 \cdots V_l) \geqq \beta$. Since by our basic assumption above, $\alpha \leqq \beta$, it follows that $\mu(U_1 \cdots U_k) \leqq \mu(V_1 \cdots V_l)$. If the additional assumption above is also valid, then a similar argument shows that $\mu(U_1 \cdots U_k) < \mu(V_1 \cdots V_l)$. □

Another useful property of averaging systems can be derived from the fact that Γ, the range of μ, is linearly ordered. While we will occasionally make tacit use of this fact, for most of our results we need only refer to the following.

LEMMA 1.2.4 (Strong Interpolation Property). *For any sequences U and V, the following are equivalent:*

(1) $\mu(U) < \mu(V)$,
(2) $\mu(U) < \mu(UV)$,
(3) $\mu(U) < \mu(VU)$,
(4) $\mu(UV) < \mu(V)$,
(5) $\mu(VU) < \mu(V)$,

and the following are equivalent:

(a) $\mu(U) = \mu(V)$,
(b) $\mu(U) = \mu(UV)$,
(c) $\mu(U) = \mu(VU)$.

Proof. That (1) implies (2)–(5) and (a) implies (b) and (c) is precisely the content of our basic axiom, the Interpolation Property. For the reverse implications, e.g. (2) → (1), (3) → (1), etc., we note that since the range of μ is linearly ordered, we must have either $\mu(U) < \mu(V)$, $\mu(U) = \mu(V)$ or $\mu(U) > \mu(V)$. Yet these latter two relationships violate conditions (2)–(5), by the Interpolation Property, and the first and third relations violate conditions (b) and (c) by the same property. □

1.3. Quasi-monotonic sequences. In this general framework we have outlined, the notion of a quasi-monotonic sequence can be given as follows.

DEFINITION. A sequence S is *quasi-increasing* (*quasi-nondecreasing*) if $\mu(U) < \mu(V)$ ($\mu(U) \leqq \mu(V)$) for all nonempty sequences, U, V such that $S = UV$. *Quasi-decreasing* and *quasi-nonincreasing* sequences are defined analogously. S is *quasi-monotonic* if it is a sequence of any of these four types.

Example. If μ is the arithmetic mean, then

2 1 3 4 is quasi-increasing,
3 4 2 1 is quasi-decreasing,
1 4 2 3 is quasi-nondecreasing,

3 2 4 1 is quasi-nonincreasing, and
1 4 3 2 is none of the above.

When μ is the lexicographic mean, the set of quasi-increasing sequences is the set of *Lyndon words* over the reals (see e.g. [Lot83]).

Two useful variants of the definition of a quasi-monotonic sequence are given in the following lemma. Here, and in several subsequent lemmas, we give only the quasi-increasing and/or quasi-nondecreasing versions, since the other cases follow by a similar argument, by simply reversing the sense of the inequalities.

LEMMA 1.3.1 (Prefix/Suffix Lemma). *Let S be a sequence. The following are equivalent*:

(i) *S is quasi-increasing* (*quasi-nondecreasing*).

(ii) $\mu(U) < \mu(S)$ $(\mu(U) \leqq \mu(S))$ *for every proper prefix U of S.*

(iii) $\mu(S) < \mu(V)$ $(\mu(S) \leqq \mu(V))$ *for every proper suffix V of S.*

Proof. This follows directly from the definition of a quasi-increasing (quasi-nondecreasing) sequence, using the Strong Interpolation Property. □

It follows that for a quasi-increasing sequence S, $\mu(U) < \mu(V)$ for any proper prefix U and proper suffix V of S, even if they overlap or are separated by some nonempty middle segment of S.

The four classes of quasi monotonic sequences have intersection properties similar to those of normal monotonic sequences, as is demonstrated in the following lemma.

DEFINITION. A sequence $S = s_1 \cdots s_n$ is *constant* if $\mu(s_i) = \mu(s_j)$ for all i and j, $1 \leqq i, j \leqq n$.

LEMMA 1.3.2 (Constant Sequence Lemma). (1) *A sequence is both quasi-nondecreasing and quasi-nonincreasing if and only if it is constant.*

(2) *A sequence is both quasi-increasing and quasi-decreasing if and only if it has only one element.*

Proof. If S is both quasi-nondecreasing and quasi-nonincreasing, then by the Prefix/Suffix Lemma above, for any proper prefix U of S, $\mu(U) \leqq \mu(S)$ and $\mu(U) \geqq \mu(S)$, i.e., $\mu(U) = \mu(S)$. Hence by the Strong Interpolation Property (repeatedly), it follows that $\mu(s_i) = \mu(S)$ for all i, $1 \leqq i \leqq n$, and thus S is constant. For the second part, if S is both quasi-decreasing and quasi-increasing then for any proper prefix U of S, $\mu(U) < \mu(S)$ and $\mu(U) > \mu(S)$; hence S has no proper prefixes, i.e., S has only one element. □

We close this introductory section by briefly examining the conditions under which quasi-monotonic sequences can be combined to form larger quasi-monotonic sequences. As above, we will restrict our attention to quasi-increasing and quasi-nondecreasing sequences. Our first lemma deals with sequences formed by concatenating quasi-nondecreasing sequences.

LEMMA 1.3.3 (Construction Lemma). *Let $S_1, \cdots, S_k$ be quasi-nondecreasing sequences, where $k > 1$, and let $S = S_1 \cdots S_k$. S is quasi-increasing* (*quasi-nondecreasing*) *if and only if $\mu(S_1 \cdots S_i) < \mu(S)$ $(\mu(S_1 \cdots S_i) \leqq \mu(S))$ for all i, $1 \leqq i < k$.*

Proof. We will prove only the "quasi-increasing part" of this result, since the other part is analogous. Further, since the "only if" implication of this part follows directly from the Prefix/Suffix Lemma, we need only verify the "if" implication.

Using the Prefix/Suffix Lemma, it suffices to show that $\mu(S_1 \cdots S_iU) < \mu(S)$ for any i, $0 \leqq i < k$, and any proper prefix U of S_{i+1}. If $i = 0$, then $\mu(S_{i+1}) = \mu(S_1) < \mu(S)$ since $k > 1$. Furthermore, since S_1 is quasi-nondecreasing, $\mu(U) \leqq \mu(S_1)$. Hence $\mu(S_1 \cdots S_iU) = \mu(U) < \mu(S)$, establishing the result. Thus we may assume that $i > 0$. Now if $\mu(U) \leqq \mu(S)$ then since $\mu(S_1 \cdots S_i) < \mu(S)$ as well, we have $\mu(S_1 \cdots S_iU) < \mu(S)$ by the Balancing Lemma. Hence we may also assume that $\mu(S) < \mu(U)$. Let V be the sequence such that $S_{i+1} = UV$. Since S_{i+1} is quasi-nondecreasing, $\mu(U) \leqq \mu(V)$. Hence

$\mu(S) < \mu(V)$. Now since either $S_{i+2} \cdots S_k$ is empty or $\mu(S) < \mu(S_{i+2} \cdots S_k)$ (by the Strong Interpolation Property), we have $\mu(S) < \mu(VS_{i+2} \cdots S_k)$ by the Balancing Lemma. Thus since $S = S_1 \cdots S_iUVS_{i+2} \cdots S_k$, by the Strong Interpolation Property, $\mu(S_1 \cdots S_iU) < \mu(S)$. □

A useful special case of the above result is the following.

COROLLARY 1.3.4 (Construction Corollary). *If $S_1, \cdots, S_k$ are quasi-nondecreasing sequences, where $k > 1$, and $\mu(S_1) < \mu(S_2) < \cdots < \mu(S_k)$ ($\mu(S_1) \leqq \mu(S_2) \leqq \cdots \leqq \mu(S_k)$) then $S_1 \cdots S_k$ is quasi-increasing (quasi-nondecreasing).*

Proof. This follows directly from the above lemma using the Balancing Lemma. □

We also consider sequences obtained by overlapping quasi-nondecreasing sequences.

LEMMA 1.3.5 (Overlap Lemma). *Let T, U, and V be sequences with U nonempty. If TU and UV are quasi-increasing (quasi-nondecreasing), then $S = TUV$ is quasi-increasing (quasi-nondecreasing).*

Proof. Again we prove only the result for the quasi-increasing case, the other case being entirely analogous.

Let $S = XY$, where X and Y are nonempty sequences. We show that $\mu(X) < \mu(Y)$. Consider three cases.

(a) $X = TL$ and $Y = RV$ for some nonempty L and R, (thus $U = LR$). Since TU is quasi-increasing, by the Prefix/Suffix Lemma, $\mu(TL) < \mu(TU) < \mu(U)$. Similarly, $\mu(U) < \mu(UV) < \mu(RV)$. Thus, $\mu(X) = \mu(TL) < \mu(RV) = \mu(Y)$.

(b) $X = L$ and $Y = RUV$ for some nonempty L, (thus $T = LR$). Following the reasoning used in (a), $\mu(L) < \mu(TU) < \mu(U)$, $\mu(RU)$. Also $\mu(U) < \mu(UV) < \mu(V)$. Thus, since $\mu(L) < \mu(RU)$ and $\mu(L) < \mu(V)$, by the Balancing Lemma

$$\mu(X) = \mu(L) < \mu(RUV) = \mu(Y).$$

(c) $X = TUL$ and $Y = R$ for some nonempty R, (thus $V = LR$). This case is a mirror image of case (b), and has a parallel proof. □

2. Trends.

2.1. The Decomposition Theorem.

In Lemma 1.3.3 (the Construction Lemma) we have considered the conditions under which quasi-increasing sequences can be concatenated to form larger quasi-increasing sequences. In this section, we consider the related problem of how we can decompose an arbitrary sequence into quasi-increasing segments. A related approach to decompositions of this type is given in [Vie78].

Since each sequence element is itself a quasi-increasing segment, we restrict our attention to segments which are quasi-increasing and of maximal length.

DEFINITION. Given a sequence $S = s_1 \cdots s_n$, a segment $U = s_i \cdots s_j$, $1 \leqq i \leqq j \leqq n$, is *maximal quasi-increasing* if U is quasi-increasing and no extension $s_h \cdots s_k$ of U, where $1 \leqq h \leqq i \leqq j \leqq k \leqq n$ and $h < i$ or $j < k$, is quasi-increasing. A maximal quasi-increasing segment is called an *upward trend*. Downward trends are defined analogously. A sequence of sequences $S_1, \cdots, S_k$ is a *decomposition of S into upward (downward) trends* if $S = S_1 \cdots S_k$ and S_i is an upward (downward) trend of S for each i, $1 \leqq i \leqq k$.

Example. If μ is the arithmetic mean, then the sequence 1 2 1 5 4 3 1 2 0 can be decomposed into upward trends as 1 2 1 5 4 3, 1 2, 0 and into downward trends as 1, 2 1, 5 4 3 1 2 0.

As in the previous section, we will state many of our results only in their quasi-increasing and/or quasi-nondecreasing versions. Unless otherwise indicated, the word *trend* indicates an upward trend and a *decomposition of S* indicates a decomposition of S into upward trends. Our main result is the following.

THEOREM 2.1.1 (Decomposition Theorem). *Any sequence S can be uniquely decomposed into upward trends, and every upward trend of S is a member of this decomposition.*

Proof. Let $S = s_1 \cdots s_n$. For each element s_i of S, find the maximal quasi-increasing segment of S which contains s_i. This segment is unique by the Overlap Lemma. Let $S_1, \cdots, S_k$ be the list of distinct segments found by successively considering elements $s_1, \cdots, s_n$. Again by the Overlap Lemma, these must form a decomposition of S into upward trends, and every trend of S must appear in this decomposition. □

When μ is the lexicographic mean, the decomposition into upward trends is known as the Lyndon factorization. When μ is the arithmetic mean, this decomposition has been called Spitzer's factorization (see [Lot83], [Spi56]).

The general relationship among the segments in the decomposition of S is outlined in the next lemma.

LEMMA 2.1.2 (Trend Mean Lemma). (1) *$S_1, \cdots, S_k$ is the decomposition of S into upward trends if and only if $S = S_1 \cdots S_k$, S_i is quasi-increasing for all i, $1 \leqq i \leqq k$, and $\mu(S_1) \geqq \mu(S_2) \geqq \cdots \geqq \mu(S_k)$.*

(2) *Let $S_1, \cdots, S_k$ be the decomposition of S into upward trends. If S is quasi-nondecreasing, then $\mu(S_1) = \mu(S_2) = \cdots = \mu(S_k) = \mu(S)$; otherwise $\mu(S_1) > \mu(S) > \mu(S_k)$.*

Proof. ad (1). If $S_1, \cdots, S_k$ is the decomposition of S, then by definition we must have $S = S_1 \cdots S_k$ and S_i quasi-increasing, $1 \leqq i \leqq k$. Further, we cannot have $\mu(S_i) < \mu(S_{i+1})$ for any i, $1 \leqq i \leqq k$, for by the Construction Corollary, this would imply that S_iS_{i+1} is quasi-increasing, contradicting the maximality of the trends. Hence $\mu(S_1) \geqq \mu(S_2) \geqq \cdots \geqq \mu(S_k)$. For the other direction, assume that these three conditions hold. Suppose that S_i is not maximal for some i, $1 \leqq i \leqq k$. Thus $S = ULS_iRV$ where U, L, R, V are sequences (either L or R may be empty, but not both) and LS_iR is maximal quasi-increasing. From the Overlap Lemma, it follows that $LS_iR = S_h \cdots S_l$ for some $1 \leqq h \leqq i \leqq l \leqq k$, where $h < i$ or $i < l$. However, then by the Prefix/Suffix Lemma, we must have $\mu(S_h) < \mu(S_l)$, a contradiction to the third condition of (1). Hence each S_i must be maximal, and thus $S_l, \cdots, S_k$ is the decomposition of S.

ad (2). If S is quasi-nondecreasing, then by the Prefix/Suffix Lemma, $\mu(S_1) \leqq \mu(S) \leqq \mu(S_k)$. Hence from part (1), $\mu(S_1) = \mu(S_2) = \cdots = \mu(S_k) = \mu(S)$. On the other hand, if S is not quasi-nondecreasing, then it cannot be the case that $\mu(S_1) = \mu(S_2) = \cdots = \mu(S_k)$ (by the Construction Corollary); hence we must have $\mu(S_1) > \mu(S_k)$ by part (1). Again from part (1), using the Balancing Lemma, it follows further that $\mu(S_1) > \mu(S) > \mu(S_k)$. □

2.2. The Collect-and-Merge Algorithm. We now turn our attention to the problem of computing the decomposition of a given sequence. We will present an algorithm that produces the decomposition of a given sequence on-line in linear time, under certain general assumptions concerning the computation of μ. We will use a model of computation in which all integers and real numbers to arbitrary precision occupy constant space, and all normal arithmetic operations, on these numbers, including addition, subtraction, multiplication and division, take constant time. This is known as the *uniform cost* RAM *model* (see e.g. [Aho74]). The key element in our algorithm is the following abstract data type.

DEFINITION. A *block* is a data type which represents specific information about an arbitrary segment of a sequence $S = s_1 \cdots s_n$. This data type supports the following functions, where b, b_1, b_2 are arbitrary blocks representing segments T, U, V respectively:

(1) *Location* (b) returns the index in S of the first letter in the segment T.

(2) *Length* (b) returns the length of T.

(3) $\mu(b)$ returns $\mu(T)$.

(4) *Merge* (b_1, b_2) returns a block b representing the segment UV if the segment V occurs immediately following U in S; otherwise it returns some special error value.

(5) *Makeblock* (T) returns a block representing the segment T.

An averaging system and its associated averaging function are *constant time merging* if for any sequence S, there is an implementation of the data type "block" for segments of S in which each block occupies constant space, each of the functions (1)–(4) defined above take constant time, and for any segment T of length 1, *makeblock* T takes constant time.

LEMMA 2.2.1. *Using the uniform cost* RAM *model, the arithmetic and gradient means are constant time merging.*

Proof. If μ is the arithmetic mean, then a block representing a segment T can be implemented as a record which consists of the index in S of the first element of T, the length of T, and the real number which gives the sum of the elements of T. It is clear that under the model of computation we are using, this data structure occupies constant space, and all of the functions associated with a block can be computed from it in the required time. For the gradient mean, we can use a similar data structure which includes both the numerator and the denominator of the fraction that defines $\mu(S)$. □

Our algorithm to find the decomposition of a sequence $S = s_1 \cdots s_n$ will create a stack of blocks $b_1, \cdots, b_k$ representing the trends $S_1, \cdots, S_k$ of this decomposition. This will be accomplished by successively computing the stacks representing the decomposition of $s_1 \cdots s_i$ for $i = 1$ to n. Thus we will need a procedure to update an existing stack of trends when a new element is added on the right end of the sequence. We give this procedure in the following general format.

THE PROCEDURE COALESCE (Q, b)

input: a sequence SB with B quasi-increasing, a stack of blocks $Q = b_1, \cdots, b_k$ (with b_k at the top) representing the decomposition $S_1, \cdots, S_k$ of S into upward trends and a block b representing the segment B. (We allow the possibility that Q is the empty stack and S is the empty sequence.)

output: a stack of blocks $Q = t_1, \cdots, t_l$ (with t_l at the top) representing the decomposition $T_1, \cdots, T_l$ of SB into upward trends.

```
begin
   while $Q$ is not empty and $\mu(\text{top}(Q)) < \mu(b)$ do
   begin
      pop the top block $b_{\text{top}}$ from $Q$;
      let $b = merge\ (b_{\text{top}}, b)$;
   end
   push $b$ onto $Q$;
end.
```

LEMMA 2.2.2. *The procedure coalesce is correct.*

Proof. If Q is empty then the while loop of *coalesce* is not executed, and it is obvious that the procedure is correct. Otherwise, we claim the while loop has the following invariant:

(1) $Q = b_1, \cdots, b_j$, for some j, $0 \leqq j \leqq k$, where the segment represented by b_h is quasi-increasing, $1 \leqq h \leqq j$, and $\mu(b_1) \geqq \mu(b_2) \geqq \cdots \geqq \mu(b_j)$;

(2) the segment represented by b is quasi-increasing, and

(3) SB is represented by $b_1 \cdots b_j b$.

It is easily verified that this invariant holds before the first execution of the loop. In this case b represents B, which is quasi-increasing by assumption, and $b_1, \cdots, b_j$ represents the decomposition of S. Thus the segment represented by b_h is quasi-increasing for all h, $1 \leqq h \leqq j$, $b_1 \cdots b_j b$ represents SB and by the Trend Mean Lemma,

$$\mu(b_1) \geqq \mu(b_2) \geqq \cdots \geqq (\mu(b_j).$$

That it is preserved by the execution of the loop body follows directly from the Construction Corollary, since b_j and b are merged only when $\mu(b_j) < \mu(b)$, and in this case the segment represented by $b_j b$ must be quasi-increasing. Since each time the loop is executed, the size of Q is reduced by one, the loop will terminate. Upon termination, in addition to conditions (1)–(3) we will have either

(4) Q is empty (i.e. $j = 0$), or

(5) $j > 0$ and $\mu(b_j) \geqq \mu(b)$.

In either case, $b_1, \cdots, b_j, b$ represents the correct decomposition for SB by the Trend Mean Lemma, and hence Q is correct following the last statement of the procedure. □

The algorithm to compute the decomposition of a sequence can now be given.

THE COLLECT-AND-MERGE ALGORITHM.
input: a nonempty sequence $S = s_1 \cdots s_n$.
output: a decomposition $S_1, \cdots, S_k$ of S into upward trends.
data structures: a stack Q of blocks.

```
begin
    let $Q$ be empty;
    for $i = 1$ to $n$ do
    begin
        let $b_{\text{new}} = makeblock\ (s_i)$;
        $coalesce\ (Q, b_{\text{new}})$;
    end;
    return a list of segments represented by the elements of $Q$
    ordered from bottom to top;
end.
```

Given the correctness of *coalesce*, it is obvious that the Collect-and-Merge Algorithm is correct. We briefly analyze the time and space requirements of this algorithm.

THEOREM 2.2.3. *Using the uniform cost* RAM *model, for any averaging system which is constant time merging, the space and time requirements of the Collect-and-Merge Algorithm are $O(n)$, where n is the length of the input sequence.*

Proof. It is clear that the space requirements are $O(n)$. To analyze the time requirements, let us for the moment discount the while loop in *coalesce*. What remains are the first and last statements of the algorithm (which take time $O(n)$), and a group of middle statements which constitute a loop which is executed n times and takes constant time for each execution. Hence the total time used is $O(n)$. In the course of all executions of the middle loop, n blocks are created by calling the function *makeblock*. Now consider the while loop we omitted. One execution of the body of this loop also takes constant time. Furthermore, every time it is executed, the number of blocks in use is reduced by one. Since n blocks are introduced during the course of execution of the entire algorithm and at least one remains when the algorithm terminates, this implies that the while loop is executed at most $n - 1$ times. Thus the total running time of the algorithm is $O(n)$. □

Before continuing to the applications of the Collect-and-Merge Algorithm, we pause to consider another use of the procedure *coalesce*. Let us assume that we have already computed the decompositions for two sequences S_1 and S_2. We can combine these decompositions into a single decomposition for the sequence S_1S_2 by the following procedure.

```
THE PROCEDURE COMBINE ($T_1$, $T_2$)
input:   two stacks of blocks $U$ and $V$ representing the decompositions of $S_1$ and
         $S_2$ into upward trends, where $U$ is ordered from bottom to top and $V$ is
         ordered from top to bottom.
output:  a sequence of blocks $Q$ representing the decomposition of $S_1S_2$.
    begin
      while $V$ is not empty and $\mu$(top ($U$)) < $\mu$(top ($V$)) do
      begin
        pop the block $b$ from the top of $V$;
        coalesce ($U$, $b$);
      end;
      return $U$ (ordered from bottom to top) concatenated with $V$
        (ordered from top to bottom);
    end.
```

By arguments similar to those given above, it is clear that this procedure is correct, and that in the worst case it takes time and space proportional to the total number of blocks in the decomposition of S_1 and S_2 for an averaging system which is constant time merging. The procedure *combine* might be used in a divide-and-conquer approach to finding decompositions. However, it is clear that since the Collect-and-Merge Algorithm is already optimal for averaging systems which are constant time merging, this approach will not be useful in this case. It may be the case though, that this procedure can be used to at least improve the expected time in certain cases when the averaging system is not constant time merging. In other cases, it appears that a more direct approach, taking advantage of special features of the averaging function μ, will yield the most efficient decomposition algorithm. An example of this is Duval's decomposition algorithm for the lexicographic mean [Duv83].

2.3. Finding convex hulls. As an example of the application of the Collect-and-Merge Algorithm, consider the problem of finding the convex hull of a set of points on the x-y plane.

Assume that we are given a sequence of points $T = (x_0, y_0), \cdots, (x_n, y_n)$, where $n > 1$, with distinct x coordinates, sorted in increasing order on the x coordinate. The convex hull of T is the smallest (minimal area) closed convex polygon that contains all of the points of T. The vertices of this polygon form a subset of T known as the set of *extremal* points of T. It is clear that the set of extremal points of T must include the first and last points of T, and these extremal points will form a degenerate polygon only in the case that all of the points of T lie on the line between the first and last points of T. The edges of the convex hull connecting the extremal points that lie on or above the line from the first to the last point of T will be called the *upper part* of the convex hull, and those connecting the extremal points that lie on or below this line will be called the *lower part*. These sets of edges are disjoint, unless the convex hull is degenerate, in which case they are identical (we generalize this observation later).

The convex hull of T can be determined as follows. From the sequence T, derive a sequence of line segments $S = (s_1, t_1) \cdots (s_n, t_n)$ where $s_i = x_i - x_{i-1}$ and $t_i = y_i - y_{i-1}$ for $1 \leqq i \leqq n$. Since T is sorted and all points have distinct x coordinates, $s_i > 0$ for all i, $1 \leqq i \leqq n$. Let μ be the gradient mean, as defined above in § 1. By Lemma 2.2.1, μ is an averaging function which is constant time merging.

Now consider an arbitrary segment $U = (x_k, y_k) \cdots (x_{k+l}, y_{k+l})$ of T, where $l > 1$, and the corresponding sequence of line segments $V = (s_{k+1}, t_{k+1}) \cdots (s_{k+l}, t_{k+l})$. By the Prefix/Suffix Lemma, V is quasi-increasing if and only if

$$\mu((s_{k+1}, t_{k+1}) \cdots (s_{k+i}, t_{k+i})) = \sum_{j=1}^{i} t_{k+j} \Big/ \sum_{j=1}^{i} s_{k+j} < \mu(V) \quad \text{for all } i, 1 \leqq i < l.$$

This is obviously equivalent to the condition that the slope of the line from (x_k, y_k) to (x_{k+i}, y_{k+i}) is less than the slope from (x_k, y_k) to (x_{k+l}, y_{k+l}) for all i, $1 \leqq i < l$, i.e., that all of the points between (x_k, y_k) and (x_{k+l}, y_{k+l}) lie below the line between these two points. Similarly, V is quasi-decreasing if and only if all intermediate points lie above the line determined by the endpoints of U. It follows easily that the decomposition of S into upward trends defines the upper part of the convex hull of T, and that the decomposition into downward trends defines the lower part. Thus using the Collect-and-Merge Algorithm, the convex hull of T can be computed on-line in linear time.

This algorithm is clearly optimal in situations where the points are given in sorted order with distinct coordinates in one dimension, e.g., in applications where the points are evaluations of a function $f(x)$ for successive values of x taken at discrete intervals. Even if two points can have the same x coordinate, we can usually get around this by perturbing the points slightly within their error range. Here, as in general, care must be taken when applying this algorithm to avoid the accumulation of round-off errors.

If the points of T are not originally given sorted on their x coordinates, then to use the Collect-and-Merge Algorithm, it requires $O(n \log n)$ time to sort them, giving a total running time $O(n \log n)$. This is the best possible time bound that can be achieved in this situation [Yao79], and there are several algorithms which achieve it, either by sorting first and then applying a hull finding procedure (which in some cases appears to be a special case of the Collect-and-Merge Algorithm, e.g., [And79]), or by using divide-and-conquer techniques (e.g. [Ben78]). Many of these latter algorithms are appealing because they run in $O(n)$ expected time for a variety of point distributions. Here it should be noted that the algorithm above, and some of the other techniques based on sorting will also run in $O(n)$ expected time if an $O(n)$ expected time sort can be used (see [Mei80] for an example of a sort which achieves this expected time for a wide class of distributions).

2.4. The Trend Boundary Theorem. The relationship between the convex hull and the corresponding decompositions given above suggests other properties of decompositions which have not yet been explored. For example, as we mentioned above, it is intuitively obvious that the set of extremal points of $T = (x_0, y_0), \cdots, (x_n, y_n)$ between (x_0, y_0) and (x_n, y_n) on which the upper part of the convex hull of T is defined is always disjoint from the set on which the lower part is defined, unless the convex hull is degenerate. This is a general phenomenon that occurs when decompositions into upward trends are compared with decompositions into downward trends. Loosely stated, our result is that internal boundaries are never shared between elements of these decompositions, unless the sequence is constant.

THEOREM 2.4.1 (Trend Boundary Theorem). *Let S be a sequence decomposed into upward trends by $I_1, \cdots, I_k$ and into downward trends by $D_1, \cdots, D_l$. If $I_1 \cdots I_r = D_1 \cdots D_s$ for any r, s, $1 \leqq r < k$ and $1 \leqq s < l$, then S is constant.*

Proof. Let $U = I_1 \cdots I_r = D_1 \cdots D_s$ and $V = I_{r+1} \cdots I_k = D_{s+1} \cdots D_l$. By the Trend Mean Lemma,

(1) $\mu(I_1) \geqq \mu(I_2) \geqq \cdots \geqq \mu(I_k)$, and

(2) $\mu(D_1) \leqq \mu(D_2) \leqq \cdots \leqq \mu(D_l)$.

Using the Balancing Lemma, from (1) we have $\mu(U) \geqq \mu(V)$ and from (2) we have $\mu(U) \leqq \mu(V)$. Thus $\mu(U) = \mu(V)$. However, now again using the Balancing Lemma, this implies that none of the inequalities in (1) or (2) above can be strict. Hence

(3) $\mu(I_n) = \mu(D_m) = \mu(S)$ for all $1 \leqq n \leqq k$ and $1 \leqq m \leqq l$.

Now since I_1 is quasi-increasing and $\mu(D_1) = \mu(I_1)$, D_1 cannot be a proper prefix of I_1. Similarly, I_1 cannot be a proper prefix of D_1; hence $I_1 = D_1$. Continuing in this manner, it follows that $k = l$ and $I_n = D_n$ for all $1 \leqq n \leqq k$, i.e., the upward and downward decompositions of S must be identical. Finally, since only a single element sequence can be both quasi-increasing and quasi-decreasing (Lemma 1.3.2), S must be constant. □

3. Cyclic conjugates.

3.1. The Rotate-and-Merge Algorithm. Many more applications of the theory of quasi-monotonic sequences can be obtained by considering the families of sequences obtained by taking all cyclic conjugates of a sequence.

DEFINITION. Given a sequence $S = s_1 \cdots s_n$, the set of *cyclic conjugates* of S is $\{S\} \cup \{s_i \cdots s_n s_1 \cdots s_{i-1} : 1 < i \leqq n\}$.

We will show that every sequence has a cyclic conjugate which is quasi-nondecreasing, and a cyclic conjugate which is quasi-nonincreasing. The key idea is given in the following.

LEMMA 3.1.1 (Trend Rotation Lemma). *Let S be decomposed into upward trends by $S_1, \cdots, S_k$. If $\mu(S_1) > \mu(S_k)$, then $T = S_2 \cdots S_k S_1$ has fewer trends than S.*

Proof. Since $S_1, \cdots, S_k$ is a decomposition of S into upward trends, each S_i, $1 \leqq i \leqq k$, is a quasi-increasing sequence. Since $\mu(S_k) < \mu(S_1)$, by the Construction Corollary, S_kS_1 is also quasi-increasing. Hence $S_2, \cdots, S_{k-1}, S_kS_1$ is a decomposition of T into $k-1$ quasi-increasing segments. Thus by the Overlap Lemma, the decomposition of T into upward trends cannot have more than $k-1$ members. □

THEOREM 3.1.2 (Cyclic Conjugate Theorem). *Every nonempty sequence S has a quasi-nondecreasing cyclic conjugate and a quasi-nonincreasing cyclic conjugate.*

Proof. From the Trend Mean Lemma, if a sequence S is not quasi-nondecreasing, the decomposition of S has more than one trend and μ of the final trend is less than μ of the first trend. Hence the Trend Rotation Lemma implies that whenever S is not quasi-nondecreasing, a cyclic conjugate T of S with fewer trends in its decomposition can be found by rotating the first trend in the decomposition of S to the back of S. By iterating this procedure, we must eventually reach a cyclic conjugate of S which has only one trend, or one with two or more trends, such that μ of the first trend is equal to μ of the last trend. In either case, this cyclic conjugate of S will be quasi-nondecreasing. A similar argument holds for quasi-nonincreasing cyclic conjugates. □

The proof of this theorem also provides us with a simple algorithm for finding a quasi-nondecreasing cyclic conjugate of an arbitrary sequence S. This algorithm is presented below. We will use the abstract data type *block* introduced in the previous section and the procedures and terminology associated with it, under the assumption that these definitions have been extended to allow us to treat a sequence S as if it was circular, so that we can merge the segment at the right end of the sequence with the segment at the left end. We also assume that the procedure *coalesce* has been extended from a stack of blocks to a queue of blocks in a natural manner, taking the back of this queue as the top of the stack.

```
THE ROTATE-AND-MERGE ALGORITHM.
input: a nonempty sequence $S = s_1 \cdots s_n$.
output: an index $j$ such that $s_j s_{j+1} \cdots s_n s_1 \cdots s_{j-1}$ is quasi-nondecreasing.
data structures: a queue $Q$ of blocks.

  begin
    apply the Collect-and-Merge Algorithm to $S$ to obtain a decomposition
      $S_1, \cdots, S_k$ of $S$ into upward trends;
    let $Q = b_1, \cdots, b_k$ be a queue of blocks representing these trends with
      front $(Q) = b_1$ and back $(Q) = b_k$;
    while $\mu$(back $(Q)$) $<$ $\mu$(front $(Q)$) do
    begin
      remove the block at the front of $Q$ and call it $b$;
      coalesce $(Q, b)$;
    end;
    return location (front $(Q)$);
  end.
```

The correctness of this algorithm is easily established, as described above. Under the assumption that μ is constant time merging, and using a uniform cost RAM model as described in the previous section, the timing analysis is also easy. It is simply an extension of the analysis of the Collect-and-Merge Algorithm given in Theorem 2.2.3. Again the critical factor is the total number of merges executed in during the course of the computation. The same reasoning of Theorem 2.2.3 applied to the Rotate-and-Merge Algorithm shows that the total number of merges executed during the first step, where the Collect-and-Merge Algorithm is called, combined with those in the remaining steps is exactly $n - 1$. Hence the Rotate-and-Merge Algorithm is $O(n)$.

3.2. Trip Around the Moon. As an application of these results, consider the following problem, known as "Trip Around the Moon" [Gra83].

You are to make one trip around the Moon in a circular path. At various points along this path, there are n fueling stations $t_1, \cdots, t_n$ with fuel amounts $f_1, \cdots, f_n$, such that the total amount of fuel available is sufficient to make one circular trip. You are not guaranteed, however, that the amount of fuel available in each station is sufficient to cover the distance to the next station. You begin at the station of your choice with an empty fuel tank. By choosing the right starting station, can you make the entire trip without running out of fuel?

The answer to this question is always yes, independently of the given configuration of fueling stations. We can demonstrate this as follows.

Let $d_1, \cdots, d_n$ be the distances between stations, where d_i is the distance between t_i and t_{i+1}, $1 \leqq i < n$ and d_n is the distance between t_n and t_1. Assume that the units chosen are such that we can travel distance d with fuel amount f if and only if $f \geqq d$. Let $S = (f_1, d_1) \cdots (f_n, d_n)$ and let μ be the gradient mean, defined in § 1. Let $T = (f_i, d_i) \cdots (f_n, d_n)(f_1, d_1) \cdots (f_{i-1}, d_{i-1})$ be a quasi-nonincreasing cyclic conjugate of S, as guaranteed by the Cyclic Conjugate Theorem. By the basic assumption of the problem, $\mu(T) = \sum_{j=1}^{n} f_j / \sum_{j=1}^{n} d_j \geqq 1$. Let U be any proper prefix of T. Since T is quasi-nonincreasing, $\mu(U) \geqq \mu(T) \geqq 1$. Hence the sum of the fuel available in the stations of U is greater than or equal to the total distance spanned by U. Since this holds for every proper prefix U of T, the trip can be made starting at station t_i.

Furthermore, since the gradient mean is constant time merging (Lemma 2.2.1), we can apply the Rotate-and-Merge Algorithm to find station t_i in time proportional to the number of fuel stations. Thus we can solve the Trip Around the Moon problem in optimal time.

Notice that in our argument, we have implicitly used the fact that the distances d_i are positive, but not the fact that the fuel amounts f_i are (presumably) positive. In fact the problem has a solution even when the f_i are allowed to be negative, since the gradient mean is still an averaging function in this case (Lemma 1.2.2). Our results may be summarized as follows.

THEOREM 3.2.1. *For any nonempty sequence* $S = (f_1, d_1) \cdots (f_n, d_n)$, *where* f_i, d_i *are real numbers,* $d_i > 0$, $1 \leqq i \leqq n$, *and* $\sum_{i=1}^{n} f_i \geqq \sum_{i=1}^{n} d_i$, *there is a cyclic conjugate* $T = (f_1', d_1') \cdots (f_n', d_n')$ *of* S *such that* $\sum_{i=1}^{j} f_i' \geqq \sum_{i=1}^{j} d_i'$ *for all* j, $1 \leqq j \leqq n$. *Such a cyclic conjugate* T *can be found in time proportional to* n, *using a uniform cost* RAM *model of computation.*

A related result appears in a recent article by Dvornicich [Dvo80] which presents some results used to derive efficient algorithms for handling data in magnetic bubble memories.

THEOREM 3.2.2 (Dvornicich). *Let* $S = s_1 \cdots s_n$ *be a sequence of real numbers such that* $\sum_{i=1}^{n} s_i \geqq 0$. *Then there is a cyclic conjugate* $T = t_1 \cdots t_n$ *of* S *such that* $\sum_{i=1}^{j} t_i \geqq 0$ *for all* j, $1 \leqq j \leqq n$.

Proof. Let T be a quasi-nonincreasing cyclic conjugate of S using the arithmetic mean. Thus $\mu(t_1 \cdots t_j) \geqq \mu(T) \geqq 0$ for all j, $1 \leqq j \leqq n$, and the result follows. □

The Dvornicich result can also be derived as a corollary to Theorem 3.2.1, or from the more general results presented in [Gra63] and [Tak67, Thm. 2, p. 1]. We have not determined if these later results can also be derived within the framework we have presented.

3.3. Unbalanced sequences and the Counting Theorem. We can obtain stronger results along the lines of the Cyclic Conjugate Theorem by undertaking a more detailed analysis of the structure of the set of cyclic conjugates of an arbitrary sequence with respect to μ. We take up this task presently.

DEFINITION. Each of the cyclic conjugates of $S = s_1 \cdots s_n$ defines the same *circular sequence* S', derived by forming the letters of S into a clockwise circular arrangement with s_1 following s_n. Segments of S' will be denoted by ranges $s_i \cdots s_j$. When $i \leqq j$, this corresponds to the standard notation. When $i > j$, $s_i \cdots s_j = s_i \cdots s_n s_1 \cdots s_j$.

Given a circular sequence S, the set of cyclic conjugates that form it can be obtained from the set of possible cuts of S.

DEFINITION. Given a circular sequence S formed from $s_1 \cdots s_n$, $C_S = \{c_1, \cdots, c_n\}$ is the set of *cuts* of S, where c_i is the cut between s_i and s_{i+1} for $1 \leqq i < n$ and c_n is the cut between s_n and s_1.

Two distinct cuts c_i and c_j in the circular sequence S formed from $s_1 \cdots s_n$ define a pair of opposing segments $s_{i+1} \cdots s_j$ and $s_{j+1} \cdots s_i$. In our basic structural result below, we express the relationship between opposing segments (with respect to μ) as a relationship between their corresponding cuts.

DEFINITION. Given a circular sequence $S = s_1 \cdots s_n$ with cuts $C_S = \{c_1, \cdots, c_n\}$, the relation $\leqq$ on C_S is defined as follows. $c_i \leqq c_j$ if and only if $i = j$, or $i \neq j$ and $\mu(s_{i+1} \cdots s_j) \leqq \mu(s_{j+1} \cdots s_i)$.

We explore the properties of the relation $\leqq$ on C_S.

DEFINITION. Given a set A and a binary relation $\leqq$ on A, $\leqq$ is a *preorder* if it is reflexive and transitive. $\leqq$ is a *linear preorder* if in addition, $a \leqq b$ or $b \leqq a$ for any a, b in A.

Our basic structural result is the following.

LEMMA 3.3.1 (Cut Order Lemma). *For any circular sequence S, $\leqq$ is a linear preorder of C_S.*

Proof. Obviously $\leqq$ is reflexive and since the range of μ is linearly ordered, for any i, j, $1 \leqq i, j \leqq n$, either $c_i \leqq c_j$ or $c_j \leqq c_i$ (or both). Now assume that $c_i \leqq c_j$ and $c_j \leqq c_k$. If i, j and k are not pairwise distinct, then it is obvious that $c_i \leqq c_k$. Otherwise, we may assume without loss of generality that $i < j < k$ or $i < k < j$. We consider only the former case. The latter case is similar. Let $X = s_{i+1} \cdots s_j$, $Y = s_{j+1} \cdots s_k$ and $Z = s_{k+1} \cdots s_i$. Since $c_i \leqq c_j$, $\mu(X) \leqq \mu(YZ)$. Since $c_j \leqq c_k$, $\mu(Y) \leqq \mu(ZX)$. Thus by the Interpolation Property (twice) we have $\mu(X) \leqq \mu(YZX) \leqq \mu(ZX)$. Hence $\mu(X) \leqq \mu(Z)$ by Strong Interpolation. Thus $\mu(ZX) \leqq \mu(Z)$, which implies that $\mu(Y) \leqq \mu(Z)$. Hence by the Balancing Lemma, $\mu(XY) \leqq \mu(Z)$, i.e., $c_i \leqq c_k$. It follows that $\leqq$ is transitive, and thus $\leqq$ is a linear preorder. □

A stronger Cyclic Conjugate Theorem will be obtained for sequences in which $\leqq$ is a linear ordering on the set of cuts. By the above lemma, these are sequences for which $\leqq$ is antisymmetric (i.e., $c_i \leqq c_j$ and $c_j \leqq c_i$ implies that $i = j$). This class of sequences can be easily characterized.

DEFINITION. A nonempty sequence S is *unbalanced* if $\mu(U) \neq \mu(S)$ for any proper prefix U of S. S is *cyclically unbalanced* if every cyclic conjugate of S is unbalanced.

LEMMA 3.3.2. *For any nonempty sequence S, $\leqq$ is a linear ordering on C_S if and only if S is cyclically unbalanced.*

Proof. This follows easily from the Cut Order Theorem, using the Strong Interpolation Property. □

To state the stronger version of the Cyclic Conjugate Theorem that holds for cyclically unbalanced sequences, we introduce the following notation.

DEFINITION. Given a nonempty sequence $S = s_1 \cdots s_n$,

$\psi(S)$ = the number of indices i, $1 \leqq i < n$, such that $\mu(s_1 \cdots s_i) \geqq \mu(S)$, and

$\psi^*(S)$ = the number of indices i, $1 \leqq i < n$, such that $\mu(s_1 \cdots s_i) > \mu(S)$.

This notation actually provides a slightly more general framework for the theory of quasi-monotonic sequences.

LEMMA 3.3.3. *For any nonempty sequence S,*
S is quasi-increasing $\Leftrightarrow \psi(S) = 0$,
S is quasi-nondecreasing $\Leftrightarrow \psi^*(S) = 0$,
S is quasi-decreasing $\Leftrightarrow \psi^*(S) = n - 1$, *and*
S is quasi-nonincreasing $\Leftrightarrow \psi(S) = n - 1$.

Proof. This is obvious. □

THEOREM 3.3.4 (Strong Cyclic Conjugate Theorem). *If $S = s_1 \cdots s_n$ is cyclically unbalanced then*

(1) $\psi(T) = \psi^*(T)$ *for every cyclic conjugate T of S and*

(2) *for every value of k, $0 \leqq k \leqq n - 1$, there is a unique cyclic conjugate T of S such that $\psi(T) = k$.*

Proof. The first part is obvious. For the second part, since S is cyclically unbalanced, $\leqq$ is a linear order on C_S by Lemma 3.3.2. Let $c_{i_1}, \cdots, c_{i_n}$ be the cuts of S listed in increasing order. For any j, $1 \leqq j \leqq n$, $c_{i_j} \leqq c_{i_m}$ for exactly $n - j$ cuts c_{i_m} distinct from c_{i_j}. Thus if T_j is the cyclic conjugate of S formed by the cut c_{i_j}, $1 \leqq j \leqq n$, then $\psi(T_j) = \psi^*(T_j) = (n-1) - (n-j) = j - 1$, $1 \leqq j \leqq n$. The result follows. □

The following corollary of this result will be useful.

THEOREM 3.3.5 (Counting Theorem). *If Ω is a set of unbalanced sequences of length n which is closed under cyclic conjugation, then for each k, $0 \leqq k \leqq n-1$, there are exactly $|\Omega|/n$ sequences S in Ω such that $\psi(S) = k$ (equivalently, $\psi^*(S) = k$).*

Proof. Obviously, by these assumptions the sequences of Ω must be cyclically unbalanced. Further, by the above result, every sequence in Ω has n distinct cyclic conjugates, all of which are in Ω. It follows that Ω can be partitioned into $m = |\Omega|/n$ classes of n sequences each, where each class is the set of all cyclic conjugates of a given sequence S. Additionally from the above result, for each possible value of ψ there is one element of each class on which ψ has this value. Thus for any particular value of ψ, ψ has this value on exactly m sequences of Ω. □

When μ is the lexicographic mean, it is clear the $\mu(X) = \mu(Y)$ if and only if $XY = YX$, i.e., if and only if there exists a nonempty sequence Z and $i, j > 0$ such that $X = Z^i$ and $Y = Z^j$ (see e.g. [Lot83]). It follows that a sequence S is unbalanced with respect to the lexicographic mean if and only if it is primitive, i.e., if and only if there exists no nonempty sequence Z and $i > 1$ such that $S = Z^i$. In this case S will be cyclically unbalanced as well, since every cyclic conjugate of a primitive sequence is primitive. Thus when μ is the lexicographic mean, Theorem 3.3.4 holds for every primitive sequence and Theorem 3.3.5 holds for every set of primitive sequences closed under cyclic conjugation. In fact this is true of any μ which has the property that $\mu(X) = \mu(Y)$ if and only if $XY = YX$.

3.4. Counting and randomly generating ordered trees. The Counting Theorem can be applied to many types of enumeration problems, and in particular, to many of those involving objects enumerated by the well-known Catalan numbers,

$$C_n = \frac{1}{n+1}\binom{2n}{n}$$

(see e.g. [Gar76], [Sin79], [Der80]). As an example, consider the following list of objects given in [Der80].

DEFINITION.

T_n is the set of rooted ordered trees with n edges (i.e. $n + 1$ nodes).

P_n is the set of *legal* sequences of n open and n closed parentheses. A parenthetical expression is called "legal" if each open parenthesis has a matching closed parenthesis.

I_n is the set *dominating* sequences $S = s_1 \cdots s_{n+1}$ of $n + 1$ nonnegative integers which sum to n, such that $\sum_{j=1}^{i} s_j \geqq i$ for all i, $1 \leqq i \leqq n$. (Because of a misprint in [Der80], we follow the definition given in [Read72] here (see also [Zak79]).)

L_n is the set of *admissible* paths from the point (0, 0) to (n, n) in an $n \times n$ lattice. All steps in a lattice path are either up or to the right; a path is "admissible" if it does not pass below the diagonal $y = x$.

B_n is the set of *full* binary trees with n internal nodes. A rooted ordered tree is "full binary" if all nodes are either of degree 0 (leaves) or of degree 2 (have exactly two successors).

Using one-to-one correspondences between these objects, by showing that the dominating sequences are enumerated by the Catalan numbers, Dershowitz and Zaks show that all of the above objects are enumerated by the Catalan numbers. We demonstrate briefly how the Counting Theorem can be applied to achieve this result. We will use the following general property of the arithmetic mean for sequences of integers.

LEMMA 3.4.1. *Let μ be the arithmetic mean and S be a sequence of integers of length n which sums to $t \neq 0$. If n and t are relatively prime then S is unbalanced.*

Proof. Let U be any proper prefix of S and let t' be the sum of U. If $\mu(U) = \mu(S)$ then $t'/|U| = t/n$, which is impossible because n and t are relatively prime and $|U| < n$. Thus S is unbalanced. □

LEMMA 3.4.2. *Let μ be the arithmetic mean and let $S = s_1 \cdots s_{n+1}$ be a sequence of $n + 1$ nonnegative integers which sums to n. Then S is a dominating sequence if and only if S is quasi-decreasing.*

Proof. If S is a dominating sequence then $\sum_{j=1}^{i} s_j/i \geqq 1 > n/(n+1) = \mu(S)$ for every i, $1 \leqq i \leqq n$. Hence S is quasi-decreasing. On the other hand, if S is quasi-decreasing, then $\sum_{j=1}^{i} s_j/i > n/(n+1) > (i-1)/i$ for every i, $1 \leqq i \leqq n$. Hence $\sum_{j=1}^{i} s_j \geqq i$ for any i, $1 \leqq i \leqq n$. Thus S is a dominating sequence. □

THEOREM 3.4.3. *$I_n = C_n$ for all $n \geqq 1$.*

Proof. Let Ω be the set of all sequences $n + 1$ nonnegative integers which sum to n and let μ be the arithmetic mean. By the Lemma 3.4.1, Ω is a set of unbalanced sequences of length $n + 1$ which is closed under cyclic conjugation. Hence by the Counting Lemma, exactly $1/(n+1)|\Omega|$ sequences from this set are quasi-decreasing. Thus we need only show that $|\Omega| = \binom{2n}{n}$ and the result will follow from Lemma 3.4.2. This latter fact is easily established by showing that every sequence $S = s_1 \cdots s_{n+1}$ in Ω can be uniquely represented by a sequence S' of n 0's and n 1's where $S' = 1^{s_1}01^{s_2}0 \cdots 01^{s_{n+1}}$, and vice versa that every such sequence represents a member of Ω. □

Since by the correspondence of [Der80], the dominating sequence associated with a given tree is simply the sequence of outdegrees of its nodes visited in preorder, we can also use these techniques to count trees whose nodes have any specific spectrum of outdegrees. If t is a tree with $n + 1$ nodes and n_i is the number of nodes with outdegree i for $1 \leqq i \leqq k$, where k is the maximal outdegree of any node, then we must have

(1) $n + 1 = n_0 + n_1 + \cdots + n_k$.

{The total number of nodes is $n+1$.}

(2) $n = n_1 + 2n_2 + \cdots + kn_k$.

{The total number of edges is n.}

THEOREM 3.4.4. *The number of rooted oriented trees with $n + 1$ nodes and n_i nodes of outdegree i for $0 \leqq i \leqq k$, where the n_i satisfy (1) and (2) above, is*

$$\frac{1}{n+1}\left[\frac{(n+1)!}{n_0!\,n_1! \cdots n_k!}\right].$$

Proof. Consider the set Ω of all sequences with n_0 0's, n_1 1's, $\cdots$, n_k k's. By (1) these sequences are of length $n + 1$ and by (2) they sum to n. Thus as in the previous theorem, exactly $1/(n+1)|\Omega|$ of these sequences are dominating sequences, i.e., represent legitimate trees. The result follows. □

One application of these results is in the generation of random trees. Using the technique from the proof of Theorem 3.4.3, we can obtain a random tree with $n + 1$ nodes by generating a random binary sequence with n 0's and n 1's, viewing it as a sequence of $n + 1$ numbers in unary separated by 0's, obtaining the quasi-increasing cyclic conjugate of this sequence of numbers (under the arithmetic mean) by the Rotate-and-Merge Algorithm, and finally interpreting the resulting sequence as the preorder traversal of a tree. To generate trees whose nodes have a specific spectrum of outdegrees, the initial sequence of $n + 1$ numbers can be chosen to reflect these constraints. This general method is well suited for efficient implementation, and thus should prove practical in situations where rapid generation of random trees is needed.

The techniques for counting trees and other objects given above are not unrelated to the specific techniques given in [Der80] and other techniques for Catalan enumerations, for example those in [Sil69], [San78] and [Sin78]. Since the literature on Catalan numbers and their relatives is extensive (a bibliography of 470 references is given in [Gou76]), no attempt has been made to cover the applications of the present theory in this area in any detail. Hence this remains an interesting area for further research.

3.5. Ballot problems. Our final application of the Counting Theorem involves a version of the classic ballot problem [Ber87].

A typical ballot problem may be described as follows. Suppose that an election is held, and candidate A receives a votes, while candidate B receives b votes. Let S be the sequence of votes as they are received, and suppose that all $\binom{a+b}{a}$ possible arrangements of S are equally likely. For a given γ, let $\Delta_\gamma(S)$ denote the number of times during the election that the ratio of votes for A to the total votes is greater than or equal to γ. For any given k, $1 \leqq k \leqq a + b$, what is the probability that $\Delta_\gamma(S) = k$?

To illustrate the application of the Counting Theorem to this type of problem, we will derive the following result from [Sri79], originally due to Takacs [Tak62]. Our version is a minor rewording of that given in [Sri79].

THEOREM 3.5.1. *If a and b are relatively prime and $\gamma = a/(a + b)$ (i.e., when γ is the final ratio of votes for A), then the probability that $\Delta_\gamma(S) = k$ is $1/(a + b)$ for all k, $1 \leqq k \leqq a + b$.*

Proof. Let S be given as a sequence of integers where each vote for A is represented by 1 and each vote for B is represented by 0. Let μ be the arithmetic mean. Thus $\mu(S) = a/(a + b) = \gamma$. Let U be a proper prefix of S of length r and let α be the number of 1's in U. The ratio of votes for A in U is greater than or equal to γ if and only if $\mu(U) = \alpha/r \geqq \gamma = \mu(S)$. Thus $\Delta_\gamma(S) = \psi(S) + 1$, since the ratio of votes for A is always greater than or equal to γ at the end of the election. Furthermore, since a and b are relatively prime, a and $a + b$ are relatively prime, and thus by Lemma 3.4.1, S is unbalanced. Since the set of all possible voting records is obviously closed under cyclic conjugation, by the Counting Theorem, all values of Δ_γ between 1 and $a + b$ are equally likely on this set, and the result follows. □

We can also obtain another related, but more general theorem of Takacs' [Tak67, Thm. 1, p. 162], using this method.

THEOREM 3.5.2 (Takacs). *Let $S = s_1 \cdots s_n$ be a sequence of integers which sums to 1. For each j, $1 \leqq j \leqq n$, there is exactly one cyclic conjugate of S for which exactly j of its partial sums are positive.*

Proof. Let μ be the arithmetic mean. Then by Lemma 3.4.1, S is cyclically unbalanced. Let $U = s_1 \cdots s_i$ be any proper prefix of S. If the partial sum $t = \sum_{k=1}^{i} s_k$ is positive, then $\mu(U) = t/|U|$ must be greater than $\mu(S) = 1/n$. On the other hand, if $\mu(U) > 1/n$, then clearly t must be positive. It follows that exactly j partial sums of S are positive if and only if $\psi^*(S) = j$. Thus the result follows from the Strong Cyclic Conjugate Theorem. □

Theorem 2.1 of [Spi56] can be derived from the Strong Cyclic Conjugate theorem in a similar manner.

4. Further research. We have already alluded to a few possible directions for further research; among them are a detailed axiomatic investigation of averaging systems and the theory of quasi-monotonic sequences, and a more extensive investigation of the time bounds for the major algorithms used in this theory, using more general assumptions concerning μ. Under the latter topic, we note that the question of parallel algorithms for finding decompositions and quasi-nondecreasing cyclic conjugates of sequences remains to be explored as well. This is an area in which we have done almost no work at this time. If good parallel algorithms are found, further applications in the area of loop or ring-structured networks (see e.g. [Dol82]) might be explored.

We also hope to use this theory to investigate certain aspects of the structure of random sequences. Using a technique similar to the one used in the proof of Theorem 2.1 in [Spi56], in certain cases we can find correspondences between the trends in the decomposition of a sequence and the cycles in a permutation of that sequence. This allows us to show, for example, that the expected number of trends (using the arithmetic mean) in a sequence of n reals randomly chosen in the interval between 0 and 1 is the same as the expected number of cycles in an arbitrary permutation of that sequence, which is ln (n).

Acknowledgments. We would like to thank Ron Graham for telling us of the "Trip Around the Moon" problem, and pointing out the applications of this theory to this problem and to the results of Takacs. We are also indebted to Joel Cohen for showing one of us a simple convex hull finding algorithm which another of us "rediscovered" in the setting of this theory.

REFERENCES

[Aho74] A. Aho, J. Hopcroft and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[And79] A. M. Andrew, *Another efficient algorithm for convex hulls in two dimensions*, Inform. Process. Lett., 9 (1979), pp. 216–219.

[Ben78] J. Bentley and M. Shamos, *Divide and conquer for linear expected time*, Inform. Process. Lett., 7 (1978), pp. 87–91.

[Ber87] J. Bertrand, *Solution d'un probleme*, Comptes Rendus des Seances de l'Academie des Sciences, Paris, 105 (1887), p. 369.

[Der80] N. Dershowitz and S. Zaks, *Enumerations of ordered trees*, Discrete Math., 31 (1980), pp. 9–28.

[Dol82] D. Dolev, M. Klawe and M. Rodeh, *An $O(n \log n)$ unidirectional distributed algorithm for extrema finding in a circle*, J. Algorithms, 3 (1982), pp. 245–260.

[Duv83] J. P. Duval, *Factorizing words over an ordered alphabet*, J. Algorithms, 4 (1983), pp. 363–381.

[Dvo80] R. Dvornicich, *On a problem of cyclic permutations of integers*, Discrete Appl. Math., 2 (1980), pp. 353–355.

[Gar76] M. Gardner, *Catalan numbers: An integer sequence that materializes in unexpected places*, Scientific American, 243 (June 1976), pp. 120–125.

[Gou76] H. W. Gould, *Research bibliography of two special number sequences*, rev. ed. 1976. (Available from author at 1239 College A., Morgantown, W.V. 26505, U.S.A.)

[Gra63] R. L. Graham, *A combinatorial theorem for partial sums*, Ann. Math. Statist., 34 (1963), pp. 1600–1602.

[Gra83] ———, personal communication.

[Lot83] M. Lothaire, *Combinatorics on words*, in Encyclopedia of Mathematics and its Applications, Vol. 17, Addison-Wesley, Reading, MA, 1983.

[Mei80] H. Meijer and S. Akl, *Design and analysis of a new hybrid sorting algorithm*, Inform. Process. Lett., 10 (1980), pp. 213–218.

[Pre79] F. P. Preparata, *An optimal real time algorithm for planar convex hulls*, Comm. Assoc. Comput. Mach., 22 (1979), pp. 402–404.

[Read72] R. C. Read, *The coding of various kinds of unlabeled trees*, in Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1972, pp. 153–182.

[San78] A. D. Sands, *On generalized Catalan numbers*, Discrete Math., 21 (1978), pp. 219–221.

[Sil69] D. M. Silberger, *Occurrences of the integer* $(2n-2)!/n!(n-1)!$, Math. Prace. Mat., 13 (1969), pp. 91–96.

[Sin78] D. Singmaster, *An elementary evaluation of the Catalan numbers*, Amer. Math. Monthly, 85 (May, 1978), pp. 366–368.

[Sin79] ———, *Some Catalan correspondences*, J. London Math. Soc. (2), 19 (1979), pp. 203–206.

[Spi56] F. Spitzer, *A combinatorial lemma and its applications to probability theory*, Trans. Amer. Math. Soc., 82 (1956), pp. 323–339.

[Sri79] R. Srinivasan, *On some results of Takacs in ballot problems*, Discrete Math., 28 (1979), pp. 213–218.

[Tak67] L. Takacs, *Combinatorial Methods in the Theory of Stochastic Processes*, John Wiley, New York, 1967.

[Tak62] ———, *Ballot problems*, Z. Wahrs., 1 (1962), pp. 154–158.

[Vie78] G. Viennot, *Algebres de Lie Libres et Monoides Libres*, Lecture Notes in Math. 691, Springer-Verlag, Berlin–New York, 1978.

[Yao79] A. Yao, *A lower bound to finding convex hulls*, J. Assoc. Comput. Mach., 28 (1981), pp. 780–789.

[Zak79] S. Zaks and D. Richards, *Generating trees and other combinatorial objects lexicographically*, SIAM J. Comput., 8 (1979), pp. 73–81.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

010

A HILL-CLIMBING ALGORITHM FOR THE CONSTRUCTION OF ONE-FACTORIZATIONS AND ROOM SQUARES*

J. H. DINITZ† AND D. R. STINSON‡

Abstract. In this paper we describe and discuss hill-climbing algorithms for the construction of one-factorizations of complete graphs, and orthogonal one-factorizations of complete graphs (i.e., Room squares).

Key words. hill-climbing, algorithm, Room squares, one-factorizations

AMS(MOS) subject classifications. 05B15, 68E10

1. Introduction. In this paper, we study hill-climbing algorithms for certain types of combinatorial designs. In the past, combinatorial designs have usually been constructed using backtracking algorithms (see [2] and [6], for example). Recently, however, hill-climbing algorithms have enjoyed some success in certain cases.

First, we briefly describe hill-climbing algorithms in a general setting. Suppose that we have some particular combinatorial optimization problem for which we want to design an algorithm. For any problem instance I, there is a set of *feasible solutions* $F(I)$; each feasible solution X has a *cost* $c(X)$. The *optimal solution* is the feasible solution X having the minimum cost. (Alternatively, we could associate a profit with each feasible solution, and ask for the feasible solution with maximum profit.)

We define a hill-climbing algorithm for a combinatorial optimization problem in terms of one or more heuristics H. Each heuristic is based on a neighbourhood system, as follows. A *neighbourhood* of X is any collection of feasible solutions $N(X)$ such that $X \in N(X)$. If, for every feasible solution X, we define a neighbourhood $N(X)$ of X, then we obtain a *neighbourhood system*. Given a neighbourhood system and a feasible solution X, the *heuristic* H nondeterministically chooses any feasible solution $Y \in N(X)$ such that $c(Y) \leqq c(X)$. If there is no such Y, then the heuristic fails, and we say that X is a *local minimum* (with respect to the heuristic H). We can define several different types of neighbourhoods and associate a different heuristic with each.

Suppose we have defined a heuristic H. Suppose also that we have some method of generating an "initial" feasible solution X. Then, the hill-climbing algorithm proceeds as follows:

```
generate initial feasible solution $X$;
while $X$ is not a local minimum do
  begin
    choose any $Y \in N(X)$ such that $c(Y) \leqq c(X)$;
    $X := Y$
  end;
```

Our hope is that the final value of X is optimal, or close to optimal.

Depending on how we define neighbourhoods, it may take a lot of time to search the entire neighbourhood for a feasible solution Y such that $c(Y) \leqq c(X)$. It is often

* Received by the editors April 28, 1986; accepted for publication November 20, 1986.

† Department of Mathematics and Statistics, University of Vermont, Burlington, Vermont 05405.

‡ Department of Computer Science, University of Manitoba, Manitoba, Canada R3T 2N2.

easier to first choose $Y \in N(X)$ nondeterministically and then check if $c(Y) \leqq c(X)$. At any given stage of the algorithm, it may be necessary to choose many such Y's before finding one with $c(Y) \leqq c(X)$.

If we take this approach, we would have to specify how many attempts we allow at each stage before we abandon the search. We would define some integer "threshold function" $f(c, I)$, which is a function of the instance I and the cost c of a feasible solution, to accomplish this. Then, we obtain the following algorithm:

```
generate initial solution $X$;
count:= 0;
repeat
  count:= count + 1;
  choose any $Y \in N(X)$;
  if $c(Y) \leqq c(X)$ then
    begin
      if $c(Y) < c(X)$ then
        count:= 0;
      $X$:= $Y$
    end
until count > $f(c(X), I)$.
```

If this approach is used, it is important to choose a suitable threshold function.

Our interest is in constructing combinatorial designs using hill-climbing algorithms. In the past, hill-climbing algorithms have been employed to successfully construct Steiner triple systems, Latin squares and strong starters. We refer the interested reader to [3], [4], [13] and [14]. Hill-climbing has been less successful in investigating other problems (see, for example, [1], [11] and [16]).

In this paper, we present new hill-climbing algorithms for some other classes of designs, namely, one-factorizations of complete graphs and Room squares. Room squares are the most "complicated" type of design for which a practical hill-climbing algorithm has been found.

2. A hill-climbing algorithm for finding one-factorizations of complete graphs. The complete graph K_n is the graph on n vertices in which every pair of points is joined by an edge. A *one-factor* of K_n is a set of $n/2$ edges that partitions the vertex set (this requires than n be even). A *one-factorization* of K_n is a set of $n - 1$ one-factors that partitions the edge set. It is well known that K_n has a one-factorization if and only if n is even. Many constructions for one-factorizations are known; a good survey is presented in [8].

In order to use a hill-climbing approach, we formulate the problem as an optimization problem. A problem instance consists only of the (even) integer n for which we want to construct the one-factorization of K_n and the set of vertices V on which K_n is defined. We will represent a one-factorization of K_n as a set $\mathbf{F}$ of pairs, each having the form $(f_i, \{x, y\})$, where $1 \leqq i \leqq n - 1$, and x and y are distinct vertices of K_n. There will be $n(n - 1)/2$ such pairs and the following properties must be satisfied:

1) Every edge $\{x, y\}$ of K_n occurs in a unique pair $(f_i, \{x, y\})$;
2) For every one-factor f_i, and for every vertex x, there is a unique pair of the form $(f_i, \{x, y\})$.

Property 1) says that every edge occurs in a unique one-factor, and property 2) says that every one-factor consists of a perfect matching.

Now, we can describe feasible solutions as being *partial* one-factorizations: in the representation above, we have a set **F** of pairs, each of which has the form $(f_i, \{x, y\})$, which satisfies the properties:

1) Every edge $\{x, y\}$ of K_n occurs in at most one pair $(f_i, \{x, y\})$;
2) For every one-factor f_i, and for every vertex x, there is at most one pair of the form $(f_i, \{x, y\})$.

We define the cost $c(\mathbf{F})$ of a feasible solution **F** to be $n(n-1)/2 - |\mathbf{F}|$, where $|\mathbf{F}|$ denotes the number of pairs in **F**. Then, it is easy to see that **F** is a one-factorization if and only if $c(\mathbf{F}) = 0$.

We must now define a heuristic. First, we construct a graph which tells us what is missing from a partial one-factorization. Given a feasible solution **F**, we define $d(\mathbf{F})$, the *defect graph* of **F**, to be the graph having vertex set $V \cup \{f_i, 1 \leqq i \leqq n-1\}$, where V is the vertex set of K_n, and having the following edges:

1) For every edge $\{x, y\}$ of K_n which does *not* occur in a pair of **F**, $\{x, y\}$ is an edge of $d(\mathbf{F})$;
2) For every f_i, $1 \leqq i \leqq n-1$, $\{f_i, x\}$ is an edge of $d(\mathbf{F})$ if and only if there is *no* pair of the form $(f_i, \{x, y\})$.

In fact, we define two heuristics, $\mathbf{H_1}$ and $\mathbf{H_2}$, based on the defect graph. We say that a vertex or one-factor is *live* if it has positive degree in the defect graph. Note that, since n is odd, if a vertex or one-factor is live, then its degree must be at least two. The heuristic $\mathbf{H_1}$ is defined as follows:

```
Given a partial one-factorization $\mathbf{F}$ with defect graph $d(\mathbf{F})$, perform the following operations:
  1) choose any live point $x$ (nondeterministically)
  2) choose any one-factor $f_i$ such that $\{x, f_i\}$ is an edge of $d(\mathbf{F})$ (nondeterministically)
  3) choose any point $y$ such that $\{x, y\}$ is an edge of $d(\mathbf{F})$ (nondeterministically)
  4) if $\{y, f_i\}$ is an edge of the defect graph, then
        replace $\mathbf{F}$ by $\mathbf{F} \cup (f_i, \{x, y\})$
     else
        there is a pair in $F$ of the form $(f_i, \{z, y\})$ $(z \neq x)$
        replace $\mathbf{F}$ by $\mathbf{F} \cup (f_i, \{x, y\}) \backslash (f_i, \{z, y\})$.
```

If we apply the heuristic $\mathbf{H_1}$, then we obtain a new feasible solution in which the cost either remains the same, or is reduced by one. Also, observe that the heuristic never "fails," since steps 1), 2) and 3) can always be performed. The heuristic $\mathbf{H_2}$ is a slight variation:

```
Given a partial one-factorization $\mathbf{F}$ with defect graph $d(\mathbf{F})$, perform the following operations:
  1) choose any one-factor $f_i$ (nondeterministically)
  2) choose any two live points $x$ and $y$ such that $\{x, f_i\}$ and $\{y, f_i\}$ are both edges of
     $d(\mathbf{F})$ (nondeterministically);
  3) if $\{x, y\}$ is an edge of the defect graph, then
        replace $\mathbf{F}$ by $\mathbf{F} \cup (f_i, \{x, y\})$
     else
        there is a pair of $\mathbf{F}$ of the form $(f_j, \{x, y\})$ $(f_j \neq f_i)$
        replace $\mathbf{F}$ by $\mathbf{F} \cup (f_i, \{x, y\}) \backslash (f_j, \{x, y\})$.
```

As was the case with $\mathbf{H_1}$, the heuristic $\mathbf{H_2}$ can always be applied, and it yields a new feasible solution with either the same cost or a cost of one unit lower.

There is no guarantee that these two heuristics are sufficient to always enable us to construct a one-factorization. It seems possible that one could reach a local minimum where no application of $\mathbf{H}_1$ or $\mathbf{H}_2$ produces a feasible solution of lower cost. However, this does not seem to happen in practice (though we cannot prove that it will never happen). In over 1000000 trials, the desired one-factorizations were always constructed. Thus, as a threshold function we can define

$$f(c,I)=\infty \quad \text{if } c>0, \qquad f(0,I)=0.$$

We can compare this hill-climbing algorithm to the algorithm to construct Steiner triple systems described in [14]. The algorithm in [14] also appears very unlikely to "fail" in practice, though it can conceivably do so [10].

We also want to note that it can be implemented so that each iteration requires only constant time. The method is similar to the hill-climbing algorithm described in [14]; so we do not describe the details here.

3. A hill-climbing algorithm for constructing Room squares. Suppose we have two one-factorizations of K_n, say $\mathbf{F} = \{f_1, \cdots, f_{n-1}\}$ and $\mathbf{G} = \{g_1, \cdots, g_{n-1}\}$. We say that $\mathbf{F}$ and $\mathbf{G}$ are *orthogonal* if any f_i and any g_j ($1 \leqq i \leqq n-1$, $1 \leqq j \leqq n-1$) contain at most one edge in common. A *Room square* of side $n-1$ is defined to be a square array R of side $n-1$, in which every cell either is empty or contains an edge of K_n, such that the filled cells in every row and every column of R form a one-factor, and such that every edge of K_n occurs in exactly one cell of R. Clearly, the rows of R will induce a one-factorization of K_n, as will the columns. Also, these two one-factorizations are orthogonal (this is equivalent to saying that no cell of R contains more than one edge of K_n). Conversely, a pair of orthogonal one-factorizations $\mathbf{F}$ and $\mathbf{G}$ give rise to a Room square in a very natural way: index the rows of a square array R by the one-factors of $\mathbf{F}$, and index the columns by the one-factors of $\mathbf{G}$, and place every edge $\{x, y\}$ in the cell (f_i, g_j), where $\{x, y\} \in f_i \cap g_j$.

Room squares were introduced by T. G. Room in 1957, though examples can be found in the literature as early as 1851 [7]. Room squares were studied extensively, but the existence question was not solved until 1975, when it was shown that there is a Room square of side n if and only if n is odd and $n \neq 3, 5$. A condensed proof is presented in Mullin and Wallis [9]. However, some of the constructions for constructing Room squares are quite complicated, and it seems worthwhile to have an algorithm for producing (many) different Room squares.

We have already noted that a Room square of side n is equivalent to a pair of orthogonal one-factorizations of order $n+1$, and that we have a practical method for constructing one-factorizations. Our strategy now is to construct a one-factorization orthogonal to a given one-factorization, thereby producing a Room square. So, suppose we have a one-factorization $\mathbf{F}$, and we wish to construct $\mathbf{G}$ orthogonal to $\mathbf{F}$. (As we construct $\mathbf{G}$, $\mathbf{F}$ remains fixed.) In terms of the Room square, we have determined the rows (say), and we are attempting to "sort out" the columns.

Let us first consider how we should modify the hill-climbing algorithm to construct a $\mathbf{G}$ orthogonal to a given $\mathbf{F}$. We will maintain the array $\mathbf{R}$, in which the rows are indexed by the one-factors of $\mathbf{F}$ and the columns are indexed by the one-factors of $\mathbf{G}$, as we proceed. At any stage of the algorithm, $R(f_i, g_j) = \{x, y\}$ if $\{x, y\} \in f_i \cap g_j$, and $R(f_i, g_j) = \emptyset$, otherwise.

We use the same two heuristics $\mathbf{H}_1$ and $\mathbf{H}_2$, as before, except now they may possibly fail, if the added constraint of orthogonality is violated. Our modified heuristic $\mathbf{H}_1$ is as follows:

```
1) choose any live point $x$ (nondeterministically)
2) choose any one-factor $g_i$ such that $\{x, g_i\}$ is an edge of $d(\mathbf{G})$ (nondeterministically)
3) choose any point $y$ such that $\{x, y\}$ is an edge of $d(\mathbf{G})$ (nondeterministically)
4) let $f_j$ be the one-factor of F which contains the edge $\{x, y\}$
5) if $R(f_j, g_i)$ is not empty then
     $H_1$ fails
   else if $\{y, g_i\}$ is an edge of the defect graph, then
     replace G by $\mathbf{G} \cup (g_i, \{x, y\})$
     define $R(f_j, g_i) := \{x, y\}$
   else
     there is a pair in G of the form $(g_i, \{z, y\})$
     replace G by $\mathbf{G} \cup (g_i, \{x, y\}) \setminus (g_i, \{z, y\})$
     define $R(f_j, g_i) := \{x, y\}$.
```

The heuristic $\mathbf{H}_2$ becomes:

```
1) choose any one-factor $g_i$ (nondeterministically)
2) choose any two live points $x$ and $y$ such that $\{x, g_i\}$ and $\{y, g_i\}$ are both edges
   of $d(\mathbf{G})$ (nondeterministically);
3) let $f_j$ be the one-factor of F which contains the edge $\{x, y\}$
4) if $R(f_j, g_i)$ is not empty then
     $H_2$ fails
   else if $\{x, y\}$ is an edge of $d(\mathbf{G})$, then
     replace G by $\mathbf{G} \cup (g_i, \{x, y\})$
     define $R(f_j, g_i) := \{x, y\}$
   else
     there is a pair in G of the form $(g_k, \{x, y\})$
     replace G by $\mathbf{G} \cup (g_i, \{x, y\}) \setminus (g_k, \{x, y\})$
     define $R(f_j, g_i) := \{x, y\}$
     define $R(f_j, g_k) := \varnothing$.
```

When we try to construct a one-factorization **G** orthogonal to a given one-factorization **F**, it often does happen that we reach dead ends. For example, consider the situation when we have a feasible solution **G** with $c(\mathbf{G}) = 1$. The defect graph $d(\mathbf{G})$ consists of a triangle, of the form $g_i\, x\, y$. No matter which heuristic we apply, we will attempt to add this triangle to **G**. However, this may violate the orthogonality constraint, as there may already be an edge $\{u, v\} \in g_i \cap f_j$, where $\{x, y\} \in f_j$. Such a situation is a local minimum (with respect to $\mathbf{H}_1$ and $\mathbf{H}_2$). Many other types of local minima can also arise, so we must define a threshold function to allow for these eventualities.

After doing some experimenting, we chose the following threshold function:

$$f(c, I) = 100 \cdot n \quad \text{if } c > 0 \quad \text{(where the instance } I \text{ consists of the graph } K_n\text{)},$$

$$f(0, I) = 0.$$

Given this choice of threshold function, we were interested in determining the probability $p(n)$ of success of the algorithm, as a function of n (the size of the instance). This probability seems impossible to estimate theoretically, so we performed a large number of experimental runs, in order to obtain an empirical result. As n varied over several values between 12 and 102, the probability $p(n)$ varied between .083 and .143, in a random fashion. The average value of p appears to be between .10 and .11, and there is no trend for $p(n)$ to increase or decrease as a function of n. We also calculated the average cost of the local minima generated. Our results are presented in Table 1.

4. Constructing Room squares with subsquares. In this section, we mention an application of our hill-climbing algorithm to an as yet unsolved problem, where we expect to be able to prove some new results. This problem concerns the existence of subsquares in Room squares. If R is a Room square of side $n - 1$, and we can find $m - 1$ rows and columns of R whose intersection, S, is a Room square in its own right (of side $m - 1$) then we say that S is a *subsquare* (of R) of side $m - 1$. Observe that we can "unplug" S and replace it by any other Room square of side $m - 1$ on the same set of vertices as S and obtain another Room square. If we unplug S from R, we refer to the resulting array as an $(n - 1, m - 1)$ *incomplete* Room square.

Observe that there can exist no Room square which contains a subsquare of side 3 or 5, but it is possible for $(s, 3)$ or $(s, 5)$ incomplete Room squares to exist. Of course, these cannot be completed.

We are interested in the following question: for what ordered pairs (s, t) does there exist an (s, t) incomplete Room square? The following necessary conditions are not difficult to prove; we refer the reader to [12] for details.

THEOREM 4.1. *If there exists an* (s, t) *incomplete Room square, where* $t \geqq 0$, *then* s *and* t *are odd positive integers,* $s \geqq 3t + 2$, *and* $(s, t) \neq (5, 1)$.

We suspect that these conditions are also sufficient, but this has not yet been proved. The best known results concerning this problem can be found in [12].

We want to modify our hill-climbing algorithm to construct (s, t) incomplete subsquares. To do this, we need to reformulate the definitions in terms of one-factorizations and modify our heuristics accordingly. This is quite straightforward.

Let $\mathbf{F} = \{f_1, \cdots, f_{n-1}\}$ be a one-factorization of K_n. Given any m vertices, Y, of the K_n, there is an induced subgraph of K_m of K_n. If there are $m - 1$ one-factors in $\mathbf{F}$

TABLE 1
Construction of Room squares.

n	# trials	# successes	Probability of success	Average time per trial*	Average cost
12	1000	126	.126	0.09	1.289
16	1000	118	.118	0.16	1.316
22	1000	97	.097	0.32	1.433
26	1000	101	.101	0.57	1.512
32	1000	99	.099	0.67	1.561
36	1000	83	.083	1.2	1.563
42	1000	103	.103	1.2	1.598
46	1000	108	.108	1.6	1.526
52	1000	98	.098	1.8	1.670
56	1000	120	.120	2.1	1.573
62	1000	89	.089	2.1	1.630
66	1000	90	.090	2.1	1.684
72	500	45	.090	3.7	1.670
76	500	50	.100	4.9	1.680
82	500	56	.112	6.2	1.680
86	500	55	.110	5.2	1.582
92	300	43	.143	5.8	1.550
96	300	35	.117	8.0	1.623
102	300	25	.083	7.3	1.707

* We implemented our algorithm in Pascal/VS and ran it on the University of Manitoba Amdahl 5850 computer.

whose intersection with Y produces a one-factorization of order m, then we say we have a *sub-one-factorization* of order m. If we remove the sub-one-factorization of order m, we obtain an *incomplete* one-factorization. We can formally define this concept as follows. We start with the graph $K_n - K_m$, where m and n are both even. A *short one-factor* is defined to be a set of $(n - m)/2$ edges that partitions the vertices not in the K_m. Then, we can define an *incomplete* (n, m) one-factorization to be a set of $n - m$ one-factors and $m - 1$ short one-factors of $K_n - K_m$, whose union contains every edge of $K_n - K_m$ exactly once.

Now, we can relate incomplete one-factorizations to incomplete Room squares. Suppose we have two incomplete (n, m) one-factorizations, say $\mathbf{F} = \{f_1, \cdots, f_{n-1}\}$ and $\mathbf{G} = \{g_1, \cdots, g_{n-1}\}$, where the short one-factors are $f_1, \cdots, f_{m-1}$ and $g_1, \cdots, g_{m-1}$. We say that $\mathbf{F}$ and $\mathbf{G}$ are *orthogonal* if any f_i and any g_j $(1 \leqq i \leqq n - 1, 1 \leqq j \leqq n - 1)$ contain at most one edge in common, and further, any f_i and any g_j $(1 \leqq i \leqq m - 1, 1 \leqq j \leqq m - 1)$ contain no edges in common. It is not difficult to see that a pair of orthogonal incomplete (n, m) one-factorizations are equivalent to an incomplete $(n - 1, m - 1)$ Room square.

Hence, it is necessary only to modify the hill-climbing algorithm for one-factorizations and Room squares to handle incomplete one-factorizations. This is very simple. When we nondeterministically generate a triple $(f_i, \{x, y\})$, say, we must first check that this triple is permissible as part of an incomplete one-factorization. That is, x and y cannot both be points in the K_m, and if f_i is a short one-factor, then neither x nor y can be in the K_m. If either of these two situations arises, then the relevant heuristic fails, and we must try again.

When constructing these incomplete designs, there is a much greater probability that a heuristic will fail, so we should adjust the threshold function accordingly, allowing more tries at each level before we give up. We have run some experiments to test how the probability of success changes with different threshold functions. For each (n, m)-incomplete Room square considered, we tried several different threshold functions, of the form $f(c, I) = K \cdot (n + 1)$, if $c > 0, f(0, I) = 0$. We tried $K = 100$, 500, 1000 and 2000, as indicated.

We obtained the following data, which we present in Table 2. Note that, for fixed n and m, the probability of success tends to decrease as the threshold is increased.

5. Applications. The main application of a hill-climbing algorithm, such as the one we describe, is to produce many different designs very quickly. In [13], a hill-climbing algorithm was used to construct 2117600 Steiner triple systems of order 19. These were then tested for isomorphism using invariants, and 2111276 of the designs were nonisomorphic.

We expect that a similar approach could successfully be used to construct large numbers of nonisomorphic one-factorizations and Room squares. Modifications of the invariants used in [13] can be used to test isomorphism in these cases, as well.

We should also mention that the time and memory requirements for these algorithms are modest enough so that they can be implemented very successfully on most microcomputers. The algorithms can very easily be animated, so an observer can watch the designs being constructed. This also makes it possible to detect when the algorithm is caught in a "vicious circle." In an interactive environment, the observer could determine when a particular run has reached a "dead end," thus obviating the need for an objective function.

The other main application of hill-climbing is to construct previously unknown designs. Since the subsquare problem for Room squares is unsolved, the hill-climbing algorithm will enable us to produce new examples of Room squares with subsquares. It

TABLE 2
Construction of (*n*, *m*) – *incomplete Room squares.*
(500 *trials of each example.*)

n	*m*	Threshold	# successes	Probability of success	Average cost
19	0	$K=100$	43	0.086	1.64
19	0	$K=500$	46	0.092	1.57
19	0	$K=1000$	52	0.104	1.49
19	0	$K=2000$	58	0.116	1.46
19	1	$K=100$	41	0.082	1.75
19	1	$K=500$	50	0.100	1.37
19	1	$K=1000$	51	0.102	1.39
19	1	$K=2000$	49	0.098	1.40
19	3	$K=100$	2	0.004	3.89
19	3	$K=500$	14	0.028	2.49
19	3	$K=1000$	14	0.028	2.57
19	3	$K=2000$	11	0.022	2.43
19	5	$K=100$	0	0.000	4.57
19	5	$K=500$	0	0.000	4.82
19	5	$K=1000$	0	0.000	4.68
19	5	$K=2000$	1	0.002	4.30
29	0	$K=100$	35	0.070	1.55
29	0	$K=500$	60	0.120	1.33
29	0	$K=1000$	58	0.126	1.35
29	0	$K=2000$	53	0.106	1.35
29	1	$K=100$	26	0.052	1.93
29	1	$K=500$	46	0.092	1.47
29	1	$K=1000$	38	0.076	1.50
29	1	$K=2000$	56	0.112	1.41
29	3	$K=100$	1	0.002	4.15
29	3	$K=500$	10	0.020	2.69
29	3	$K=1000$	15	0.030	2.32
29	3	$K=2000$	31	0.062	1.85
29	5	$K=100$	0	0.000	5.85
29	5	$K=500$	0	0.000	5.13
29	5	$K=1000$	1	0.002	4.39
29	5	$K=2000$	3	0.006	4.93
39	0	$K=100$	33	0.066	1.67
39	0	$K=500$	61	0.122	1.44
39	0	$K=1000$	54	0.108	1.46
39	0	$K=2000$	46	0.092	1.39
39	3	$K=100$	3	0.006	3.95
39	3	$K=500$	29	0.058	2.10
39	3	$K=1000$	39	0.078	1.74
39	3	$K=2000$	42	0.084	1.70

should not be difficult to find an example of any particular order. Hopefully, recursive techniques will then lead to a complete solution of this problem.

For other applications of hill-climbing algorithms in obtaining new results in design theory, we refer the reader to [3], [4], [14] and [15].

REFERENCES

[1] C. J. Colbourn and E. Mendelsohn, *Kotzig factorizations: existence and computational results*, Ann. Discrete Math., 12 (1982), pp. 65–78.

[2] M. J. Colbourn, *Algorithmic aspects of combinatorial designs: a survey*, Ann. Discrete Math., 26 (1985), pp. 67–136.

[3] J. H. Dinitz and D. R. Stinson, *A note on Howell designs of odd side*, Utilitas Math., 18 (1980), pp. 207–216.

[4] ———, *A fast algorithm for finding strong starters*, this Journal, 2 (1981), pp. 50–56.

[5] ———, *The spectrum of Room cubes*, European J. Combin., 2 (1981), pp. 221–230.

[6] P. B. Gibbons, *Computing techniques for the construction and analysis of block designs*, Ph.D. thesis, University of Toronto, Toronto, Ontario, Canada, 1976.

[7] T. P. Kirkman, *Note on an unanswered prize question*, Cambridge and Dublin Math. J., 5 (1850), pp. 255–262.

[8] E. Mendelsohn and A. Rosa, *One-factorizations of the complete graph—A survey*, J. Graph Theory, 9 (1985), pp. 43–65.

[9] R. C. Mullin and W. D. Wallis, *The existence of Room squares*, Aequationes Math., 13 (1975), pp. 1–7.

[10] K. T. Phelps, Private communication.

[11] D. P. Shaver, *Construction of (v, k, λ) configurations using a non-enumerative search technique*, Ph.D. thesis, Syracuse University, Syracuse, NY, 1973.

[12] D. R. Stinson, *Room squares and subsquares*, Proc. Combinatorial Mathematics X, Adelaide, Australia, 1982, pp. 86–95.

[13] D. R. Stinson and H. Ferch, 2000000 *Steiner triple systems of order* 19, Math. Comp., 44 (1985), pp. 533–535.

[14] D. R. Stinson, *Hill-climbing algorithms for the construction of combinatorial designs*, Ann. Discrete Math., 26 (1985), pp. 321–334.

[15] D. R. Stinson and S. A. Vanstone, *A few more balanced Room squares*, J. Austral. Math. Soc. Ser. A, 39 (1985), pp. 344–352.

[16] M. Tompa, *Hill-climbing: a feasible search technique for the construction of combinatorial configurations*, M.Sc. thesis, University of Toronto, Toronto, Ontario, Canada, 1975.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

011

ON GOSSIPING WITH FAULTY TELEPHONE LINES*

RAMSEY W. HADDAD†‡, SHAIBAL ROY†§ AND ALEJANDRO A. SCHÄFFER†¶

Abstract. In the well-known gossip problem, each of n gossips initially has a unique piece of information. The gossips can make a sequence of two-party telephone calls in which the two participants exchange every piece of information they have at the time of the call. The problem is to determine a minimum length sequence of telephone calls such that, by the end, everyone knows everyone else's information. We consider Berman and Hawrylycz's variation on this problem [this Journal, 7 (1986), pp. 13–17]. They introduce the additional feature that as many as k of the calls may fail in the sense that no information is exchanged, where k is a second parameter of the problem. We improve upon their upper bound on the minimum number of calls needed. This disproves a conjecture in the same paper. We also briefly consider the parallel complexity of this problem.

Key words. gossip problem, telephone problem, fault-tolerant communication

AMS(MOS) subject classifications. 05C38, 94C15

1. Introduction. In the well-known problem of gossips and telephones, first proposed by Boyd, n gossips each have a unique piece of information, and they seek to find out everyone's information after a sequence of telephone calls. During each telephone call, the two participants exchange every piece of information they have at the time of the call. The goal is to minimize the total number of telephone calls. Baker and Shostak [1], Bumby [3] and Hajnal, Milner and Szemerédi [4] and Tijdeman [8] have shown that $2n - 4$ calls are necessary and sufficient for $n \geqq 4$. Many restrictions and variations of the problem have also been considered (cf. the bibliography of [9], and more recent papers [6], [7]).

In this paper we consider a variation on Boyd's problem proposed by Berman and Hawrylycz [2]. Berman and Hawrylycz introduce the additional feature that as many as k of the calls may fail in the sense that no information is exchanged, where k is a second parameter of the problem. The sequence of calls attempted is *static*, i.e., the gossips cannot attempt different calls depending on which ones have failed previously. Berman and Hawrylycz seek bounds on $\tau(n, k)$, the number of telephone calls needed to ensure that all n gossips possess all n pieces of information even if some arbitrary subset of k calls all fail. They show that

$$\left\lceil \left(\frac{k+4}{2}\right)(n-1) \right\rceil - 2\lceil \sqrt{n} \rceil + 1 \leqq \tau(n,k) \leqq \left\lfloor \left(k+\frac{3}{2}\right)(n-1) \right\rfloor \quad \text{for } k \leqq n-2,$$

$$\left\lceil \left(\frac{k+3}{2}\right)n \right\rceil - 2\lceil \sqrt{n} \rceil \leqq \tau(n,k) \leqq \left\lfloor \left(k+\frac{3}{2}\right)(n-1) \right\rfloor \quad \text{for } k \geqq n-2$$

and conjecture that the upper bound is almost tight—more precisely that

$$\tau(n,k) = \left(k+\frac{3}{2}\right)n - c$$

* Received by the editor June 2, 1986; accepted for publication November 20, 1986.

† Computer Science Department, Stanford University, Stanford, California 94305.

‡ The work of this author was supported by National Science Foundation grant DCR-8308109.

§ The work of this author was supported by National Science Foundation grant DCR-8351757, by Office of Naval Research contract N000 14-85-0731, and by a grant from the AT&T Foundation.

¶ The work of this author was supported by a National Science Foundation Graduate Fellowship.

where c is bounded as n grows without bound and k stays fixed. We disprove their conjecture for all but small fixed values of k by exhibiting a family of calling sequences, indexed on $m \leqq \log_2 n$ that prove the following theorem.

THEOREM 1.

$$\tau(n,k)\leqq\left(\frac{k}{2}+2m\right)\left((n-1)+\frac{n-1}{2^m-1}+2^m\right).$$

If $k \gg \log n$, then by choosing $m = \lceil(\log_2 n)/2\rceil$ the ratio of the number of calls in our construction to the lower bound proved in [2] tends to 1 in the limit, since

$$\begin{aligned}\tau(n,k)&\leqq\left(\frac{k}{2}+2\left\lceil\frac{\log_2 n}{2}\right\rceil\right)\left((n-1)+\frac{n-1}{2^{\lceil(\log_2 n)/2\rceil}-1}+2^{\lceil(\log_2 n)/2\rceil}\right)\\ &\leqq\left(\frac{k}{2}+2+\log_2 n\right)(n+3\sqrt{n})\\ &=\frac{nk}{2}+O(k\sqrt{n}+n\log_2 n).\end{aligned}$$

This is an improvement on the upper bound in [2] when

$$(k/2+2+\log_2 n)<(k+3/2);$$

i.e., when $k > 1 + 2\log_2 n$. We can improve on their bounds for smaller k by choosing $m = \lceil\log_2((k+5)/4)\rceil$ to get

$$\tau(n,k)\leqq\frac{nk}{2}+\left(\frac{2k+10}{k+1}\log_2(k+5)-\frac{10}{k+1}\right)(n-1)+O(k^2).$$

For fixed, small values of k the lower order terms of the above are not insignificant when compared to the $nk/2$ term. These bounds can be improved further by making m appropriately small. This increases the factor in front of the nk term, but the savings in the low order terms make up for it. By using $m = 2$ we get

$$\tau(n,k)\leqq\frac{2}{3}(k+8)(n-1)+2k+16,$$

whereas $m = 4$ gives

$$\tau(n,k)\leqq\frac{8}{15}(k+16)(n-1)+8k+128.$$

The former is an improvement on the bound of [2] when $(2/3)(k+8) < (k+3/2)$, that is when $k \geqq 12$, and the latter when $(8/15)(k+16) < (k+3/2)$, or $k \geqq 16$.

The rest of the paper is organized as follows. Section 2 defines the relevant concepts and proves a lemma used in the following sections. Section 3 essentially proves Theorem 1 for the special case when the number of vertices is a power of 2. Section 4 generalizes this result to an arbitrary number of vertices. Section 5 discusses the parallel complexity of this problem.

2. Preliminaries. We model the sequence of calls by a simple graph on n vertices with multiple numeric labels on each edge. The vertices represent the different gossips. Each label on an edge represents an attempted telephone call; the numbers denote the temporal order in which the calls are made. We call two paths *edge disjoint* if they share no edges and *label disjoint* if they share no edge-label pairs. Two label disjoint paths may share an edge if the labels associated with that edge on the two paths are different (i.e., the calls along that edge are made at different times). In order for vertex v to receive the information possessed by u, there must be a path of successful calls with strictly ascending labels from u to v. Since one cannot know a priori which k calls will fail, any calling sequence will be successful if and only if for *every* pair of distinct vertices (u, v) there are $k + 1$ label disjoint ascending paths from u to v.

As in the construction of Berman and Hawrylycz there will be sets of calls that can (among themselves) be made in any order. Within each such set, all calls will have the same label. Since we do not specify the order within such a set, only one member of the set can appear in any ascending path.

Our constructions repeat a fixed sequence of calls over and over. Let a *phase* be an ordered sequence $(E_1, E_2, \cdots, E_m)$ of subsets of E which denotes a sequence of calls with the meaning that one call is made along each edge in E_i after all calls in E_{i-1}. Implicitly, each edge of E_i in phase r is labeled with the ordered pair (r, i), where the phase number r is considered more significant than the subphase number i. (Alternately, this could be viewed as each edge being labeled $i + m(r - 1)$.) We specify no order for the calls within any particular E_i.

Our first lemma suggests that regardless of the structure of the graph of calls, it is a good strategy to find many short edge disjoint paths connecting each pair of vertices in the graph and to make calls repeatedly along those paths. Because the paths are edge disjoint, no sequence of calls along one of the paths ever interferes with a sequence of calls along another path. If we repeatedly try the calls along each of the paths the cost of a failed call is not too great because we will try to make another call along the same edge in the next phase. Moreover, calls in the current phase that succeed may still be used in conjunction with calls from other phases to establish label disjoint ascending paths between other vertices.

LEMMA 1. *If r phases establish p edge disjoint ascending paths from u to v, then $r + s$ phases establish $(s + 1)p$ label disjoint ascending paths from u to v.*

Proof. Let $P_1, P_2, \cdots, P_p$ be the p edge disjoint ascending paths from u to v. Any P_i consists of a sequence of edges labeled with a phase number and a subphase number. Let $P_i(j)$ be the path obtained by using the same sequence of edges as in P_i, but with the phase numbers of the edges incremented by j.

Since P_i is edge disjoint from the other P_g's (with $g \neq i$), none of the $P_i(j)$'s share any edge with $P_g(h)$ for any h. Because of the different phase numbers, no $P_i(j)$ shares an edge-label pair with $P_i(h)$, provided $h \neq j$. All the $P_i(j)$, for $0 \leqq j \leqq s$, are completed by the end of phase $r + s$. Hence, we have $(s + 1)p$ label disjoint ascending paths from u to v. □

COROLLARY 1. *If r phases give us p edge disjoint ascending paths between every pair of vertices, then we can guarantee at least $k + 1$ label disjoint ascending paths between every pair of vertices in $\lceil (k + 1)/p \rceil + r - 1$ phases.*

3. A construction for $n = 2^m$.

We begin with a construction that establishes $\tau(n, k) \leqq nk/2 + O(n \log n)$ assuming $n = 2^m$ for some integer m. Even with this severe restriction on n, this bound does not improve on the previous bound unless k grows much faster than $\log n$. However, in the next section we generalize the construction to improve the bound of Berman and Hawrylycz for arbitrary n and all but small fixed values of k.

A *hypercube* of order 2^m (or an m-cube) is the undirected graph $G = (V, E)$ where $V = \{0, 1\}^m$ and $E = \{(x0y, x1y) : x0y \in V\}$. Let $E^{(j)} := \{(x0y, x1y) \in E : |y| = j - 1\}$ for $1 \leqq j \leqq m$ denote the set of edges in *dimension* j of the m-cube (note that $E = \bigcup_j E^{(j)}$). Each vertex is incident to exactly one edge in $E^{(j)}$ for each j. Associated with any path $(u_0, u_1, \cdots, u_r) \in V^{r+1}$ in an m-cube is a unique *dimension sequence*

$(d_1, d_2, \cdots, d_r) \in \{1, 2, \cdots, m\}^r$ such that $(u_{i-1}, u_i) \in E^{(d_i)}$. Given a particular vertex, any dimension sequence identifies a unique walk starting from that vertex in the m-cube; the walk may or may not be a path. One trivial sufficient condition for such a walk to be a path, which we use implicitly in the proof of Lemma 2, is that all dimensions appear at most once in the dimension sequence. From now on, we shall denote paths in an m-cube by the starting vertex and the dimension sequence. We call a path in an m-cube and its dimension sequence, $d_1, d_2, \cdots, d_r$, *p-fold* if there are at most p indices $1 \leqq r < i$ such that $d_i \geqq d_{i+1}$.

In this construction two gossips participate in the same phone call only if they correspond to adjacent vertices in the hypercube, with the possible exception of some final special rounds of $2n - 4$ calls.

LEMMA 2. *There are m (internally) vertex disjoint* 1*-fold paths between any pair* (u, v) *of distinct vertices in an m-cube.*

Proof. Let $u[j]$ denote the jth bit of u. Let $P := \{j : u[j] = v[j]\}$, and

$$Q := \{1, 2, \cdots, m\} \backslash P.$$

In any dimension sequence specifying a path from u to v, any element of P must appear an even number of times, while any element of Q must appear an odd number of times. Let $R := (q_1, q_2, \cdots, q_{|Q|})$ denote the sequence listing the elements of Q in ascending order. Let $R(i)$ be the cyclic rotation $(q_i, q_{i+1}, \cdots, q_{|Q|}, q_1, \cdots, q_{i-1})$ for $1 \leqq i \leqq |Q|$.

The first $|Q|$ paths from u to v are given by the dimension sequences

$$\{R(i) : 1 \leqq i \leqq |Q|\}.$$

To see that these paths are internally vertex disjoint observe that if $i \neq j$, then any proper nonempty prefixes of $R(i)$ and of $R(j)$ do *not* contain exactly the same elements.

For each $p \in P$, identify the unique i_p such that the sequence $pR(i_p)p$ is 1-fold. If $p < q_1$ or $p > q_{|Q|}$ then i_p is 1, otherwise i_p is the unique integer such that $q_{i_p - 1} < p < q_{i_p}$. The remaining $|P|$ paths are given by the dimension sequences $\{pR(i_p)p : p \in P\}$. The respective first dimensions of these $|P|$ paths do not occur in any of the other dimension sequences, so all the paths are internally vertex disjoint.

Finally, note that the $|Q| + |P| = m$ paths are each of length at most $m + 1$. □

COROLLARY 2. *After* 2 *phases of* $(E^{(1)}, E^{(2)}, \cdots, E^{(m)})$, *there are m edge-disjoint ascending paths from any vertex u to any other vertex v.*

LEMMA 3. *If* $n = 2^m$ *then*

$$\tau(n, k) \leqq \min\left(\left(\left\lceil \frac{k+1}{m} \right\rceil + 1\right)\frac{mn}{2}, \left(\left\lfloor \frac{k+1}{m} \right\rfloor + 1\right)\frac{mn}{2} + (k+1 \bmod m)(2n-4)\right).$$

Proof. From Corollaries 1 and 2, we can see that repeating $\lceil (k+1)/m \rceil + 1$ phases of $(E^{(1)}, E^{(2)}, \cdots, E^{(m)})$ will suffice for k-failure-safe total communication. The number of calls in this case is $(\lceil (k+1)/m \rceil + 1)(mn/2)$.

We note that because of the ceiling function, it might be preferable for certain values of n and k to replace the last phase with j iterations of a calling sequence requiring $2n - 4$ calls such as that given in [1] that create j (final) ascending paths between any pair of vertices. In this scheme, the total number of calls is

$$\left(\left\lceil \frac{k+1-j}{m} \right\rceil + 1\right)\frac{mn}{2} + j(2n-4).$$

For any interesting value of n, a choice of $j > (k + 1 \bmod m)$ will never be optimal. □

COROLLARY 3. *If* $n = 2^m$, *then* $\tau(n, k) \leqq nk/2 + O(n \log n)$.

4. A generalized construction. The previous construction, using a single hypercube, works well for $n = 2^m$ and $k \gg \log n$. We can generalize this construction to achieve improved bounds for arbitrary n and small k.

Define an (h, m)-hypercube system as follows: Take h hypercubes, each having 2^m vertices (we will choose m later), select one vertex from each hypercube (say 0^m), and coalesce these h vertices into a single vertex. The resulting graph has $h(2^m - 1) + 1$ vertices. All telephone calls take place along hypercube edges.

LEMMA 4. *We can achieve k-failure-safe total communication for an (h, m)-hypercube system with $(\lceil (k+1)/m \rceil + 3)h2^m m/2$ calls.*

Proof. Let E_i be the union of edges in the $E^{(i)}$ of all the hypercubes. A single phase of $(E_1, E_2, \cdots, E_m)$ will use $h2^m m/2$ calls.

It follows from Lemma 2 that after 4 phases there will be m edge disjoint ascending paths between every pair of vertices in this graph. It takes 2 phases to get all the information to the central vertex belonging to all the hypercubes, and 2 more phases to disseminate the information to all the other vertices. By Corollary 2, there will be $k + 1$ label disjoint ascending paths between every pair of distinct vertices after $\lceil (k+1)/m \rceil + 3$ phases. Thus, the total number of calls needed to achieve k-failure-safe total communication for this system is at most $(\lceil (k+1)/m \rceil + 3)h2^m m/2$. □

Choosing $h = \lceil (n-1)/(2^m - 1) \rceil$ yields a graph with

$$\tilde{n} = \lceil (n-1)/(2^m-1) \rceil (2^m - 1) + 1 \geqq n$$

vertices. If the above inequality is strict, then there are fewer than $2^m - 1$ vertices left for the last hypercube. In this case some real vertices assigned to that hypercube may simulate the actions of several other positions in the hypercube that are not occupied by a real vertex. In fact, in any construction one can make any single real vertex simulate many unfilled positions in order to supply necessary symmetry; therefore, $\tau(n, k) \leqq \tau(\tilde{n}, k)$. Note that such simulation may require the creation of multiple copies of an edge in the graph to ensure the existence of enough edge disjoint paths because some of the requisite paths in the full hypercube (or other construction) may pass through positions that are not occupied by real vertices. Thus, the number of calls is

$$\begin{aligned}\tau(n,k) \leqq \tau(\tilde{n},k) &\leqq \left(\left\lceil \frac{k+1}{m} \right\rceil + 3\right)\left\lceil \frac{n-1}{2^m-1} \right\rceil 2^m \frac{m}{2} \\ &\leqq \left(\frac{k}{m} + 4\right)\left(\frac{n-1}{2^m-1} + 1\right) 2^m \frac{m}{2} \\ &= \left(\frac{k}{2} + 2m\right)\left((n-1)\frac{2^m}{2^m-1} + 2^m\right) \\ &= \left(\frac{k}{2} + 2m\right)\left((n-1) + \frac{n-1}{2^m-1} + 2^m\right),\end{aligned}$$

as claimed in Theorem 1. □

5. Parallel complexity. With slight modifications, our schemes can also yield calling sequences that are efficient in the amount of *time* they take to complete which can be viewed as the *parallel* complexity of the problem. Assuming that each call takes unit time and that each gossip participates in at most one call at a time, Knödel [5] showed that $\log_2 n$ time is required even in the absence of failures ($k = 0$). Since each gossip must make at least k telephone calls and can attempt only one call per unit time, max $(k, \log_2 n)$ is a lower bound on the parallel complexity.

As noted in § 2, our scheme specifies no order on calls with the same label, and hence, these can be made simultaneously. The only thing that might prevent this is if two edges with the same label share a vertex. In our § 3 construction, when n is a power of 2 (and when we do not use any final sequences of $2n - 4$ calls), no two edges with the same label share a vertex. Thus, since the calling sequence outlined in the proof of Lemma 3 uses at most $(\lceil (k+1)/m \rceil + 1)m \leqq k + 2 \log_2 n$ distinct labels, it requires exactly that much time.

However, in the scheme of § 4, when n is not a power of 2, the central vertex that belongs to all the hypercubes has many incident edges with the same label. Thus the scheme is not efficient in this parallel sense. If k is small this seems unavoidable because one cannot afford to reduce the degree at the cost of increasing the diameter of the graph. For $k \gg \log n$, however, we can modify our construction to improve its parallel performance without too much degradation to its total number of calls.

For our new construction let $m = \lceil (\log_2 n)/2 \rceil$, $h = \lceil n/2^m \rceil$ (and hence, $h \leqq 2^m$). We have a *central hypercube* with 2^m vertices. Also we try to have $h \leqq 2^m$ *outer hypercubes* with 2^m vertices each. Each outer hypercube shares one vertex with the central hypercube—a different vertex for each outer hypercube. If there are not enough vertices to fill the outer hypercubes, then we leave two of them incomplete, so that each one has at least 2^{m-1} vertices. The total number of vertices in this graph if all the outer hypercubes are complete is

$$\tilde{n} = (h+1)2^m - h \geqq \lceil n/2^m \rceil 2^m \geqq n.$$

A phase consists of calls along all of the edges in the outer hypercube, followed by calls along all of the edges in the central hypercube; the calls are made in increasing order of dimension in both cases. Thus each phase involves $2m$ labels, and at most $(h+1)2^m m/2$ edges. Four phases and the outer part of the fifth phase establish m edge disjoint ascending paths between any pair of verticies. Thus, we need $\lceil (k+1)/m \rceil + 4$ phases, and

$$(\lceil (k+1)/m \rceil + 4)2m \leqq 2k + 5 \log_2 n + 10$$

distinct labels and units of time are needed to establish k-failure-safe total communication if the outer hypercubes are complete. If two of the outer hypercubes have to simulate vertices, we may have to slow down the outer phases of the calling sequence by as much as a factor of two.

6. Conclusion. We have exhibited constructions that improve the bounds for the number of calls needed to achieve k-failure-safe total communication for various ranges of k. The bound

$$\tau(n, k) \leqq \frac{8}{15} nk + O(n + k)$$

is an improvement for almost all k. If k is sufficiently large, then the bound

$$\tau(n, k) \leqq \frac{nk}{2} + O(k\sqrt{n} + n \log_2 n)$$

is even closer to the lower bound. For the special case of $n = 2^m$, we can get a further small improvement to

$$\tau(n, k) \leqq \frac{nk}{2} + O(n \log_2 n).$$

We have also considered the parallel complexity of this problem and shown that the parallel time used by our scheme is within a small multiplicative factor of the optimal.

Acknowledgments. Professors Ernst Mayr and Jeffrey Ullman made helpful comments on an earlier draft.

REFERENCES

[1] B. Baker and R. Shostak, *Gossips and telephones*, Discrete Math., 2 (1972), pp. 191–193.

[2] K. A. Berman and M. Hawrylycz, *Telephone problems with failures*, this Journal, 7 (1986), pp. 13–17.

[3] R. T. Bumby, *A problem with telephones*, this Journal, 2 (1981), pp. 13–18.

[4] A. Hajnal, E. C. Milner and E. Szemerédi, *A cure for the telephone disease*, Canad. Math. Bull., 15 (1972), pp. 447–450.

[5] W. Knödel, *New gossips and telephones*, Discrete Math., 13 (1975), p. 95.

[6] A. L. Liestman and D. Richards, *Toward optimal gossiping schemes with conference calls*, Discrete Appl. Math., 7 (1984), pp. 183–189.

[7] Á. Seress, *Gossiping old ladies*, Discrete Math., 46 (1983), pp. 75–81.

[8] R. Tijdeman, *On a telephone problem*, Nieuw Arch. Wisk., 19 (1971), pp. 188–192.

[9] D. B. West, *A class of solutions to the gossip problem, part* I, Discrete Math., 39 (1982), pp. 307–326.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

COMPUTING A SPARSE BASIS FOR THE NULL SPACE*

JOHN R. GILBERT† AND MICHAEL T. HEATH‡

Abstract. We present algorithms for computing a sparse basis for the null space of a sparse underdetermined matrix. We describe several possible computational strategies, both combinatorial and noncombinatorial in nature, and we compare their effectiveness for several test problems.

Key words. null basis, null space, sparse matrix, bipartite graph, matching

AMS(MOS) subject classifications. 65F50, 68R10

1. Introduction. Let A be an $m \times n$ matrix of rank r. (Without loss of generality, we will assume throughout that $r \leqq m \leqq n$.) If B is an $n \times (n-r)$ matrix of rank $n-r$ such that

$$AB = 0,$$

then the columns of B form a basis for the $(n-r)$-dimensional null space of A. For brevity, we will refer to such a matrix B as a *null basis*. We will refer to the individual columns of B as *null vectors*, each of which corresponds to a set of columns of A whose linear combination is equal to zero. Obviously such a matrix B is not unique, not only in the relatively trivial sense of different possible scalings and column permutations, but also in the sense that there may be structurally distinct null bases for the same A (i.e., involving different combinations of columns of A).

The matrix B may be represented either explicitly, by computing its elements, or implicitly, as a product of transformations. An implicit representation suffices for some purposes; for others an explicit representation is necessary. In this paper we consider the problem of computing B explicitly.

More specifically, if the matrix A is sparse, we wish to compute a suitably sparse null basis B. It is difficult to define precisely what we mean by a "suitably" sparse null basis. For one thing, there may be no such sparse B. For example, the matrix

$$[I, e],$$

where I is the identity matrix and e is the column vector all of whose components are equal to 1, is quite sparse but has no explicit sparse representation for its one-dimensional null space. Moreover, even if a sparse null basis exists, the problem of computing a sparsest representation for it has been shown to be NP-hard [3]. As in many sparse matrix computations, we will therefore content ourselves with developing heuristic computational strategies that find a "good" sparse null basis, though not necessarily the sparsest possible.

The sparse null basis problem has at least one important feature that distinguishes it from most other sparse matrix problems. The analysis of most sparse matrix problems is simplified by ignoring any zeros that might be created through exact cancellations as a result of some arithmetic operation on nonzeros (see, e.g., [6, p. 27]). In computing a sparse null basis, however, we are specifically seeking nontrivial linear combinations of

* Received by the editors March 24, 1986; accepted for publication December 24, 1986. This work was supported by the Applied Mathematical Sciences Research Program of the Office of Energy Research of the U.S. Department of Energy under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. Publication of this report was partially supported by the National Science Foundation under grant DCR-8451385.

† Department of Computer Science, Cornell University, Ithaca, New York 14853.

‡ Mathematical Sciences Section, Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831.

nonzeros that give a zero result (i.e., arithmetic cancellation). For this reason we will necessarily employ numerical techniques along with some standard combinatorial methods, such as bipartite matching.

Before proceeding with a discussion of the algorithms we have developed, we will first give some applications that justify our interest in the sparse null basis problem and review other work on it. We then state our basic strategy for computing a sparse null basis and explore in detail several possible variations. The results of extensive empirical testing and some final observations conclude the paper.

2. Applications. There are numerous applications in which a null basis is important. The fundamental fact on which most of these applications is based is that the general solution of an underdetermined system of linear equations

$$Ax = b \tag{2.1}$$

can be expressed as

$$x = \hat{x} + By \tag{2.2}$$

for some vector y, where $\hat{x}$ is any particular solution to the system and B is a null basis. In constrained optimization problems, for example, if a set of linear (or linearized) equality constraints is expressed in the form (2.1), then every feasible point can be expressed in the form (2.2), thereby allowing the constrained problem to be solved by means of an unconstrained problem in the variable y. See [2, pp. 99–104] or [7, pp. 155–163] for further discussion of such null space methods in optimization.

The specific application that motivated our own interest in the null basis problem is the force method of structural analysis. Here A is the equilibrium matrix of a structure and b is a vector of applied loads, so that (2.1) expresses a constraint on the system force vector x, which is to be determined. (See [10] and references therein for further details of the discussion to follow.) The locality of connections within the structure causes the matrix A to be quite sparse.

Minimizing the potential energy requires that x minimize the quadratic form

$$\tfrac{1}{2}x^TDx$$

subject to the constraint (2.1), where the $n \times n$, symmetric, block-diagonal matrix D is the element flexibility matrix of the structure. Using (2.2), we see that y must satisfy the symmetric linear system

$$B^TDBy = -B^TD\hat{x}. \tag{2.3}$$

In this context the null basis B is called the self-stress matrix. Thus, having computed a particular solution $\hat{x}$ and the redundant force vector y, the desired system force vector x is given by (2.2). Even if B is sparse, B^TDB may be dense; in this case, B^TDB would be used in factored form to solve (2.3) by an iterative technique such as conjugate gradients.

One of the principal virtues of the force method is that it separates the computation into two somewhat independent phases:

(1) Compute a null basis B and a particular solution $\hat{x}$.

(2) Solve the linear system (2.3).

The importance of this separation becomes apparent when solving a sequence of problems having a fixed layout but differing material properties, such as multiple redesign problems or nonlinear elastic analysis. In such cases the matrix A is fixed, but the matrix D changes from problem to problem. Thus the first phase need be done only once for the entire sequence of problems, and only the second phase is repeated for each problem. A further implication is that it may be worth considerable effort to produce a sparse B, since this one-time cost will be amortized over the whole sequence of problems.

Another important conclusion we can draw from the various uses of the null basis is that it should be as well conditioned as possible (i.e., the columns of B should not be nearly linearly dependent numerically). For example, a poorly conditioned B would make the conditioning of the linear system (2.3) extremely poor and might therefore yield a highly inaccurate solution. For numerical purposes, an orthogonal null basis would be highly desirable, but in many cases this goal would conflict too greatly with sparsity considerations.

3. Methods for computing a null basis. Many mathematical programming algorithms use a variable-reduction technique to compute a null basis B (see, e.g., [7, p. 163]). Assume for the moment that A has full row rank m, and let A be partitioned so that

$$AP=[A_1,A_2]$$

where A_1 is $m \times m$ and nonsingular, and P is a permutation matrix that may be required in order to ensure that A_1 is nonsingular. We may then take

$$B=P\begin{bmatrix}-A_1^{-1}A_2\\ I\end{bmatrix}. \tag{3.1}$$

A permutation P that yields a structurally nonsingular A_1 can be chosen purely symbolically (see, e.g., [5]), but this says nothing about the possible numerical conditioning of A_1 and the resulting B.

In order to control numerical conditioning, numerical pivoting must be employed. Several such methods have been proposed based on various matrix factorizations, including *LU*, *QR*, *LQ*, *SVD*, and Gauss–Jordan elimination (see [10] for a survey). For example, *QR* factorization with column pivoting (see, e.g., [8, p. 165]) yields

$$AP=Q[R_1,R_2]$$

where P is again a permutation matrix, and R_1 is an upper triangular matrix of order m. We may now take

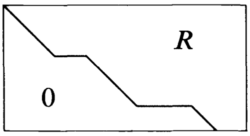
$$B=P\begin{bmatrix}-R_1^{-1}R_2\\ I\end{bmatrix}. \tag{3.2}$$

We note that if the permutation matrix P were the same in both cases, then the null bases given by (3.1) and (3.2) would be the same. Thus, the *QR* approach can be viewed simply as a means of choosing a permutation P on numerical grounds. Of course, numerical considerations may be at odds with sparsity considerations, and a compromise may have to be made between the two. In any case, with either (3.1) or (3.2) there may be a great deal of intermediate fill during the computation. Moreover, forcing B to contain an embedded identity matrix may restrict us to a considerably less sparse null basis than might otherwise be possible.

When A is banded, a method for computing a banded null basis B has been developed by Topcu [15] and Kaneko, Lawo and Thierauf [11]. Their method is based on *LU* factorization and is called, for reasons that will become obvious, the "turnback" method. Heath, Plemmons and Ward [10] extended and adapted this method for use with *QR* factorization; see also Berry et al. [1]. Our algorithms, described in §§ 4 and 5, were motivated by turnback; thus we describe this method in some detail below.

Write $A = (a_1, a_2, \cdots, a_n)$ by columns. A *start column* is a column a_s such that the ranks of $(a_1, a_2, \cdots, a_{s-1})$ and $(a_1, a_2, \cdots, a_s)$ are equal. Equivalently, a_s is a start column if it is linearly dependent on lower-numbered columns. The coefficients of this linear dependency give a null vector whose highest-numbered nonzero is in position s. It is easy to see that the number of start columns is $n - r$, the dimension of the null space of A.

The start columns can be found by doing a QR factorization of A, using orthogonal transformations to annihilate the subdiagonal nonzeros. Suppose that in carrying out the QR factorization we do not perform column interchanges but simply skip over any columns that are already zero (or numerically negligible) on and below the diagonal. The result will be a factorization of the following form:

$A = Q$  .

The start columns are the columns where the upper triangular structure jogs to the right; that is, a_s is a start column if the highest nonzero position in column s of R is no larger than the highest nonzero position in earlier columns of R.

Turnback finds one null vector for each start column a_s by "turning back" from column s to find the smallest k for which columns $a_s, a_{s-1}, \cdots, a_{s-k}$ are linearly dependent. The null vector has nonzeros only in positions $s - k$ through s. Thus if k is small for most of the start columns, then the null basis will have a small profile. Note that turnback operates on A, not R. The initial QR factorization of A is used only to determine the start columns, and is then discarded.

As described above, the null vector that turnback finds from start column a_s may not actually be nonzero in position s. Therefore, turnback needs to have some way to guarantee that its null vectors are linearly independent. Heath, Plemmons and Ward accomplish this by forbidding the leftmost column of the dependency for each null vector from participating in any later dependencies. Thus, if the null vector for start column a_s has its first nonzero in position $s - k$, every null vector for a start column to the right of a_s will be zero in position $s - k$.

4. Overview of the algorithms. The four algorithms we compare in this paper fit the following framework, which is based on turnback.

```
Preorder the columns of $A$;
Perform $QR$ factorization of $A$ to get start column numbers $s_1, s_2, \cdots, s_{n-r}$;
for $j := 1$ to $n - r$ do
    Find a null vector whose highest nonzero position is $s_j$
```

The initial QR factorization is done by the George–Heath algorithm [9], which uses sparse data structures and Givens rotations to avoid intermediate fill. As in turnback, the factorization is used only to find the start columns, and is then discarded. (In the context of solving an underdetermined system (2.1), (2.2), the initial QR factorization can be used to obtain the particular solution $\hat{x}$.) Preordering the columns of A may be necessary to make the initial QR factorization sparse. We experimented with several preordering strategies, as described in § 6.

Each start column is the rightmost member of some dependent set of columns. Thus, each start column corresponds to a null vector whose highest-numbered nonzero is in that column. Each such null vector is found independently. The null basis therefore contains an embedded triangular matrix, which is the rows of B corresponding to start columns of A.

The algorithm maintains a set of *active columns*, initially containing only the current start column a_s. It adds lower-numbered columns to the active set, one at a time. If a lower-numbered column is dependent on some active columns not including the start column, that column is not added to the set. When the active set becomes linearly dependent, its columns correspond to the nonzero positions of the desired null vector.

The *active rows* are the rows of A in which some active column is nonzero. The *active submatrix* is the matrix of active rows and columns. Thus, the algorithm keeps adding columns to the active submatrix until it becomes deficient in column rank. In order to produce a sparse null vector, we want the active submatrix to grow as little as possible before a dependency is found.

The algorithm for finding one null vector is summarized in the following pseudocode.

```
Active Columns := $\{a_s\}$;
repeat
    Choose an inactive column $a_c$, $c < s$;
    if $a_c$ is independent of Active Columns $- \{a_s\}$
        then Active Columns := Active Columns $+ \{a_c\}$
until Active Columns is linearly dependent
```

The algorithm determines linear dependence or independence by maintaining a QR factorization of the active submatrix. The factorization is stored as a single matrix, each column of which contains a column of R above the diagonal and a Householder transformation below the diagonal. It is updated as follows. Suppose there are k active columns, and a new column is being considered as a potential active column. The first k Householder transformations are applied to the new column. If the result has any nonzeros below position k, a new Householder transformation is computed to zero the new column below position $k + 1$ (thus updating the QR factorization) and the new column becomes active.

If, on the other hand, the result is zero below position k, then the new column is dependent on the other active columns. Either the dependency includes the start column, in which case the desired null vector has been found; or the dependency excludes the start column, in which case the new column does not become active and the QR factorization is not updated. Once a dependency has been found, the numerical values of the nonzero entries in the corresponding null vector are computed by back substitution with the triangular matrix R.

This procedure guarantees that the active columns are always linearly independent, so when we find the null vector it will be nonzero in position s as desired. Notice that this part of the algorithm is purely numerical; we use no information about the nonzero structure except in the definition of an active row as one in which an active column is nonzero.

The size of the active submatrix is crucial to the efficiency of the algorithm in three ways. First, its QR factorization dominates the space required by the algorithm. Second, updating this QR factorization dominates the total time required. Third, the number of columns in the active submatrix is at least as large as the number of nonzeros in the current null vector, so small active submatrices will lead to a sparse null basis. We want somehow to select columns for the active submatrix in a way that will keep the active submatrix small. The next section considers several strategies for selecting columns.

The *QR* factorization of the active submatrix tends to have a banded structure, regardless of the strategy used to select active columns. The lower band structure arises because the Householder transformation in a given column affects only rows that were active when that column was added to the submatrix, so that column of the factorization is zero in all subsequent active rows. The upper band structure arises because a newly active column may not have any nonzero rows in common with a particular previous active column, in which case that previous column's Householder transformation does not affect the new column. The factorization of the active submatrix, therefore, can be stored in a profile data structure: For each column, record the row numbers of the first and last nonzeros in the column, and store only the values in that range of rows. Our experience is that this usually reduces the storage requirement for the active submatrix to between 10–25% of the storage required for a dense matrix.

5. Details of column selection strategy. The heart of the algorithm is the strategy for choosing columns to add to the active submatrix. Since finding the sparsest null vector of a matrix is NP-hard [3], we do not hope to find the best possible choice of columns. Rather, we consider several heuristics.

5.1. Closest column next (Turnback). The simplest strategy is to choose columns in right-to-left order from the start column a_s. This is the "turnback" strategy described in § 3, with a minor difference in the way it avoids finding linearly dependent null vectors. The turnback algorithm in [10] never adds to the active submatrix the column corresponding to the lowest-numbered nonzero component of any earlier null vector; our implementation may add such a column, but it never adds a column that would create a dependency that does not include a_s.

Turnback performs well when the columns in a dependent set are close together in A, which happens when A is banded. Turnback tries to minimize the bandwidth of the current null vector, so it tries to produce a banded null basis.

Our experience is that turnback usually produces a basis with a sparse band, even when the band is fairly narrow. Therefore, a general sparse data structure may be more compact than a band or profile data structure. Our implementation stores both A and the null basis B by columns in the general sparse data structure used in Sparspak [6].

5.2. Cheapest column next. Turnback tries to find null vectors with small bandwidth by choosing columns close to the start column. In a general sparse setting we want to choose columns on grounds of sparsity rather than bandwidth. The next algorithm assigns each column a *cost* that measures the growth it would cause in the active submatrix; then the algorithm chooses the cheapest column.

Let n be the number of columns in A. The cost of column a_j is defined to be as follows:

$$\begin{aligned} cost(a_j) = {} & (\text{number of nonzeros in inactive rows of } a_j) \\ & -(\text{number of nonzeros in active rows of } a_j)/n \\ & -j/n^2. \end{aligned}$$

This definition makes a column cheaper if it adds fewer rows to the active submatrix. The null vector is sparse if the final active submatrix has few columns. The submatrix has few columns if it has few rows, since the null vector is complete when the submatrix becomes deficient in column rank.

In case of a tie in the number of nonzeros in inactive rows, we make a column cheaper if it has more nonzeros in active rows. The heuristic reason for this is that we hope to encounter numerical cancellation that will make the active submatrix deficient in column rank while it still has more rows than columns. Our experience is that such cancellation is more likely if the active submatrix is denser.

If ties in cost still remain, we make a column cheaper if it is farther to the right, that is, closer to the start column. All else being equal, this tries to minimize bandwidth. Our experience is that this tiebreaking rule usually makes little difference in the sparsity of the basis, but on some banded problems it helps significantly. Cheapest-column-next with cost defined only by this tiebreaking rule is the same as turnback.

5.3. Choosing a column by matching. A disadvantage of cheapest-column-next is that it can add a "useless" column that contains nonzeros in some inactive rows, but is linearly dependent on the active columns if we consider only the active rows. We can avoid most of the useless columns by using the combinatorial structure of the matrix. In this section we describe two versions of a heuristic that uses matchings in bipartite graphs to guide the search for a good column. The Appendix contains the necessary definitions and lemmas from bipartite matching theory.

In combinatorial terms, the matrix A is a bipartite graph whose two disjoint sets of vertices are its rows and its columns, and whose edges are its nonzeros. The start column a_s is a vertex of A. The active submatrix is the subgraph containing the active columns, the active rows (which are the vertices adjacent to active columns), and the edges between them. The active columns less a_s are always independent, so by Lemma 3 there is a matching that covers all the active columns except a_s. The new column to be added will increase the size of the matching by one. The algorithm searches for a column to add by following alternating paths. We give details below, followed by a proof that the algorithm will find the desired null vector.

Though we use matchings and alternating paths to guide the algorithm, we still use the numerical QR factorization of the active submatrix to decide when sets of columns are dependent. This lets us avoid any no-cancellation assumptions, and it lets us find null vectors that could not be predicted from the structure alone of A. It means that we must be careful to distinguish numerical and structural notions both in the algorithm and in its correctness proof: "dependent" and "independent" are numerical; "matching," "path," and "cover" are structural.

The algorithm. Given a start column a_s, this algorithm finds a null vector whose highest nonzero position is s, if there is such a null vector. In the algorithm, C is the set of active columns. The active rows are all those rows in which some active column is nonzero.

```
$C := \{a_s\}$;
Start with the empty matching;
**repeat**
    Find an alternating path from some uncovered active row, number $r$, to some
        inactive column $a_c$ to the left of $a_s$ that is independent of $C - \{a_s\}$;
    Alternate along the path, increasing the size of the matching by one
        and covering row $r$ and column $c$;
    $C := C + \{a_c\}$;
**until** either $a_s$ is dependent on $C - \{a_s\}$
    or no such alternating path exists;
```

```
if $a_s$ is dependent on $C - \{a_s\}$
    then null vector := coefficients of the dependency
    else report "no such null vector"
```

An invariant that is true at the beginning and end of the main loop is: The columns in $C - \{a_s\}$ are independent, and the matching covers all columns in $C - \{a_s\}$ and only active rows. See Fig. 1 for a sketch. (The active columns and rows may not actually be contiguous in the matrix.)

If there is no numerical cancellation, so that the rank of every matrix involved is equal to its maximum matching size, then the algorithm will find a null vector when the active columns first become more numerous than the active rows. Then there will be exactly one more active column than active row, and the matching will cover all the active rows. If there is cancellation, the algorithm may find a null vector when there are more active rows than columns. This will be a null vector whose nonzero structure could not have been predicted from the nonzero structure of A.

The nonzero positions in the null vector correspond to columns of A that are reachable by alternating paths from the start column. Alex Pothen's thesis [3] shows that (ignoring numerical cancellation) a null vector can always be found by finding a maximum matching, choosing an uncovered column as a starting column, and following all alternating paths from that column.

Correctness of the algorithm.

THEOREM. *If there is a null vector whose highest nonzero position is s, this algorithm stops with a_s dependent on $C - \{a_s\}$; otherwise it stops with a_s independent of $C - \{a_s\}$.*

Proof. Each iteration of the loop makes C larger, so the algorithm must stop eventually. If there is no null vector, a_s is independent of all the earlier columns, so it is independent of $C - \{a_s\}$. Thus we need only prove that if the null vector exists, then the algorithm will not stop early; that is, if the null vector exists and a_s is independent of $C - \{a_s\}$, then there is an alternating path from some uncovered active row to some inactive column to the left of a_s.

Assume that the desired null vector exists. Then a_s is a linear combination of the columns to the left of a_s. It is not a linear combination of $C - \{a_s\}$, so there is some

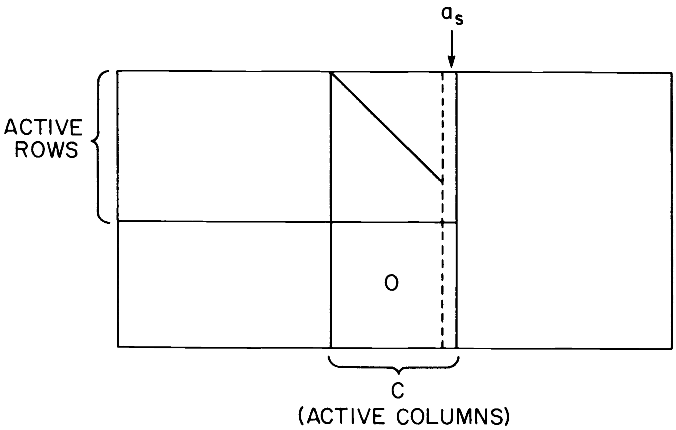


FIG. 1. *Computing one null vector. Columns in $C - \{a_s\}$ are independent and matched to active rows. Active rows and columns may not be contiguous.*

$c < s$ such that a_c is independent of $C - \{a_s\}$, even considering only the active rows. Then $C - \{a_s\} + \{a_c\}$ is independent, even considering only the active rows. Then Lemma 3 says that $C - \{a_s\} + \{a_c\}$ has a matching that covers all its columns and covers only active rows.

The current matching is thus not a maximum matching on columns $C - \{a_s\} + \{a_c\}$ and the active rows. Therefore, by Lemma 2, there is an alternating path from some uncovered active row to some uncovered column. The only uncovered column in $C - \{a_s\} + \{a_c\}$ is the inactive column a_c. □

Finding an alternating path. The algorithm maintains a queue of uncovered active rows. At each iteration, it takes a row from the head of the queue and searches for an alternating path to an inactive column. If no such path exists, it proceeds to the next row on the queue. When it chooses a new column, it adds any newly active rows to the tail of the queue.

There are two versions of the search for an alternating path from a particular uncovered active row. The "DFS matching" version performs a depth-first search through alternating paths from the row, visiting every inactive column that can be reached by such a path. It chooses the cheapest of those columns according to the cost criterion of § 5.2. The "greedy matching" first looks for an inactive column that can cover the uncovered active row, and chooses the cheapest such column if there is one. If there is no such column, it performs a depth-first search. Thus it first tries to find an alternating path of length one, and spends the time to search all alternating paths only if that fails.

The greedy algorithm is based on Duff's code MC21A for finding a nonzero diagonal of a matrix [5]. MC21A finds a matching by repeatedly finding alternating paths, and the greedy heuristic speeds up MC21A very significantly in practice. We expected the DFS version to find sparser null bases than the greedy version, but to take longer. However, in our experiments, the DFS version was usually better than the greedy version in running time and storage as well as in sparsity of the null basis. Presumably this is because it was more successful in keeping the active submatrix small; the time spent doing depth-first searches was saved in updating the QR factorization.

6. Experimental results. We experimented with the four algorithms described in § 5: turnback, cheapest column next, greedy matching, and DFS matching. Table 1 describes nine sample problems, from various sources, that we used for testing.

Our code gives the option of preordering the columns of A before beginning the null basis computation. Preordering may be necessary to keep the initial QR factorization,

TABLE 1

Description of test problems. ADLITTLE, SHARE1B, *and* MIXED1 *are linear programming problems. The remaining problems are from structural analysis* [1], [11].

Problem	Rows	Columns	Nonzeros
FRAME2D	27	45	93
PLANE	40	80	168
ADLITTLE	57	97	465
PLATE	59	144	364
FRAME3D	72	144	304
WHEEL	96	120	420
WRENCH	112	216	490
SHARE1B	118	225	1182
MIXED1	171	320	906

which is used to find start columns, sparse; for details see [9]. We found little overall correlation between preordering methods and null basis density, though some problems did show a marked preference for one ordering or another. For the results in Tables 2 through 4 we used, for each algorithm and each problem, the preordering that gave the sparsest null basis found for that algorithm on that problem.

For all but one of the problems we tried three orderings: the original order in which the matrix was presented, reverse Cuthill–McKee, and nested dissection. (The last two were applied to the structure of A^TA.) One matrix, WHEEL, was presented in four different orders; we tried all four, for a total of six in all. It should be noted that for most of the problems the original ordering had already been carefully chosen to reflect certain structural characteristics. In general, one of the automated orderings would be necessary in order to make the initial sparse QR factorization feasible.

Our experimental code is written in Fortran 77 and was run on a Vax 780 (with floating point accelerator), under Berkeley 4.2 Unix. The current version of the code does not take advantage of the profile structure of the active submatrix; it stores and manipulates the submatrix as a dense array, but measures the storage actually required by the profile. Storing only the profile would almost always save 75–95% of the storage needed for the active submatrix. This suggests that in most cases the size of the profile is only $O(m)$, instead of the worst-possible $O(m^2)$. The profile itself usually contains less than 10% zero elements.

Table 2 reports the raw results. For each problem and algorithm, the "best preorder" is the one that gave the sparsest null basis. The running times should be taken with a large grain of salt, because much of the time is spent manipulating zeros outside the profile of the active submatrix. An improved implementation that took advantage of this profile structure would presumably give much smaller times for all four algorithms. The "dense size of the active submatrix" is the product of the largest number of columns and the largest number of rows in an active submatrix during the computation. Except for $O(m + n)$ overhead, it is the intermediate storage required if the algorithm uses a dense array for the active submatrix. The "profile of the active submatrix" is the size of the largest profile of an active submatrix during the computation. Except for $O(m + n)$ overhead, it is the intermediate storage required if the algorithm uses a profile data structure for the active submatrix.

Table 3 normalizes the results in Table 2: The first column gives the ratio of the density of the particular null basis to that of the sparsest null basis found for that problem. Thus, for example, the turnback null basis for FRAME3D had 1.43 times as many nonzeros as the DFS matching null basis, which was the sparsest one found. The other columns give time, submatrix size, and submatrix profile, similarly normalized.

Table 4 reports, for each algorithm, the average over all nine problems of the normalized basis density, running time, submatrix size, and submatrix profile.

7. Conclusions. Finding a sparse null basis is a problem that is partly combinatorial, partly numerical. We have experimented with algorithms that use the combinatorial structure of the matrix to guide a search for sparse null vectors, but use numerical computation to decide linear dependence. They range in combinatorial sophistication from turnback (which uses none of the structure of the matrix), through cheapest-column-next (which uses the nonzero counts of the rows and columns), to the depth-first search matching algorithm (which uses matchings and alternating paths in the bipartite graph of the matrix).

The results in § 6 are somewhat mixed, but we can draw some rough conclusions. The DFS matching algorithm looks promising. It has a small but consistent advantage

TABLE 2
Performance measures. The "best" preorder is the one that produces the sparsest null basis.

Problem	Algorithm	Best preorder	Nonzeros in null basis	Running time	Dense size of active submatrix	Profile of active submatrix
FRAME2D	Turnback	none	76	2.30	288	27
	Cheapest Column	none	76	1.12	182	25
	Greedy Matching	none	87	0.72	110	38
	DFS Matching	none	76	0.63	72	25
PLANE	Turnback	none	166	7.05	506	82
	Cheapest Column	none	166	5.25	506	74
	Greedy Matching	RCM	183	3.45	600	204
	DFS Matching	none	177	4.88	650	119
ADLITTLE	Turnback	RCM	391	28.02	2550	597
	Cheapest Column	ND	385	17.47	1680	408
	Greedy Matching	RCM	387	8.78	1089	507
	DFS Matching	ND	367	10.50	1680	371
PLATE	Turnback	RCM	326	13.02	812	104
	Cheapest Column	RCM	310	6.65	182	68
	Greedy Matching	RCM	313	3.97	210	79
	DFS Matching	RCM	311	4.12	182	66
FRAME3D	Turnback	none	452	66.12	2064	104
	Cheapest Column	none	338	16.10	1089	48
	Greedy Matching	none	369	2.95	288	83
	DFS Matching	none	317	2.95	168	41
WHEEL	Turnback	order 3	503	17.32	1560	236
	Cheapest Column	order 2	516	83.35	9312	190
	Greedy Matching	RCM	625	16.83	2256	720
	DFS Matching	order 3	488	10.45	756	273
WRENCH	Turnback	none	544	58.38	5112	451
	Cheapest Column	none	549	57.18	3782	270
	Greedy Matching	none	590	30.52	3192	580
	DFS Matching	none	518	26.98	3782	266
SHARE1B	Turnback	none	1531	773.15	13570	8717
	Cheapest Column	ND	1567	365.90	9310	3189
	Greedy Matching	RCM	1604	98.45	8742	2770
	DFS Matching	RCM	1363	103.68	8160	1784
MIXED1	Turnback	none	1518	288.87	10506	849
	Cheapest Column	RCM	1323	584.68	15500	734
	Greedy Matching	none	1161	70.73	4356	661
	DFS Matching	none	1101	60.80	3540	697

in sparsity; indeed, for only one problem (PLANE) did it fail to come within 1% of the sparsest basis we could find. On the other hand, all the algorithms found pretty good null bases; the worst basis of the four was rarely more than 25% denser than the best. (This is counting only actual nonzeros; a band-oriented approach like the original turnback algorithms might also require storage of many zero entries in the null basis.) The differences in running time and storage were greater: DFS matching usually required substantially less space than the nonmatching methods, measuring space either by dense size or profile size of the active submatrix. Running times seem to correlate well with active submatrix size.

TABLE 3
Normalized performance measures.

Problem	Algorithm	Best preorder	Nonzeros in null basis	Running time	Dense size of active submatrix	Profile of active submatrix
FRAME2D	Turnback	none	1.00	3.65	4.00	1.08
	Cheapest Column	none	1.00	1.78	2.53	1.00
	Greedy Matching	none	1.14	1.14	1.53	1.52
	DFS Matching	none	1.00	1.00	1.00	1.00
PLANE	Turnback	none	1.00	2.04	1.00	1.11
	Cheapest Column	none	1.00	1.52	1.00	1.00
	Greedy Matching	RCM	1.10	1.00	1.19	2.76
	DFS Matching	none	1.07	1.41	1.28	1.61
ADLITTLE	Turnback	RCM	1.07	3.19	2.34	1.61
	Cheapest Column	ND	1.05	1.99	1.54	1.10
	Greedy Matching	RCM	1.05	1.00	1.00	1.37
	DFS Matching	ND	1.00	1.20	1.54	1.00
PLATE	Turnback	RCM	1.05	3.28	4.46	1.58
	Cheapest Column	RCM	1.00	1.68	1.00	1.03
	Greedy Matching	RCM	1.01	1.00	1.15	1.20
	DFS Matching	RCM	1.00	1.04	1.00	1.00
FRAME3D	Turnback	none	1.43	22.41	12.29	2.54
	Cheapest Column	none	1.07	5.46	6.48	1.17
	Greedy Matching	none	1.16	1.00	1.71	2.02
	DFS Matching	none	1.00	1.00	1.00	1.00
WHEEL	Turnback	order 3	1.03	1.66	2.06	1.24
	Cheapest Column	order 2	1.06	7.98	12.32	1.00
	Greedy Matching	RCM	1.28	1.61	2.98	3.79
	DFS Matching	order 3	1.00	1.00	1.00	1.44
WRENCH	Turnback	none	1.05	2.16	1.60	1.70
	Cheapest Column	none	1.06	2.12	1.18	1.02
	Greedy Matching	none	1.14	1.13	1.00	2.18
	DFS Matching	none	1.00	1.00	1.18	1.00
SHARE1B	Turnback	none	1.12	7.85	1.66	4.89
	Cheapest Column	ND	1.15	3.72	1.14	1.79
	Greedy Matching	RCM	1.18	1.00	1.07	1.55
	DFS Matching	RCM	1.00	1.05	1.00	1.00
MIXED1	Turnback	none	1.38	4.75	2.97	1.28
	Cheapest Column	RCM	1.20	9.62	4.37	1.11
	Greedy Matching	none	1.05	1.16	1.23	1.00
	DFS Matching	none	1.00	1.00	1.00	1.05

TABLE 4
Normalized performance measures, averaged over all nine problems.

Algorithm	Nonzeros in null basis	Running time	Dense size of active submatrix	Profile of active submatrix
Turnback	1.13	5.67	3.60	1.89
Cheapest Column	1.07	3.99	3.51	1.14
Greedy Matching	1.12	1.12	1.43	1.93
DFS Matching	1.01	1.08	1.11	1.12

Cheapest-column-next generally did better than turnback but worse than DFS matching. On the whole, we conclude that more combinatorial sophistication seems to help, both in sparsity of the null basis and in effort to find it.

All these algorithms are limited by the storage needed for the active submatrix. The *QR* factorization of the active submatrix has a profile structure, but it is dense within the profile, and we know of no way to reduce the size of the profile while adding columns to the submatrix in unpredictable order. The other storage bottleneck is the initial *QR* factorization of A. Here we use Heath's technique of withholding any dense rows if necessary [9]. This approach may lead to a few extra "spurious" start columns, but these are detected correctly by the subsequent null vector algorithm and do not affect the ultimate null basis (although they may incur extra computation). We do not have detailed statistics on the relative seriousness of these two bottlenecks.

We made some tests of the numerical quality of the computed null basis B. We estimated $\|AB\|/\|A\|\,\|B\|$ to see how nearly orthogonal A and B were, and the answer was always near machine precision. We estimated the condition number of B for six of the problems (all but MIXED1, ADLITTLE, and SHARE1B). The answer was almost always reasonably small, but on PLANE and WHEEL there were a few bases with conditions as high as 10^8. The ill-conditioned bases do not seem to correlate with choice of algorithm or choice of preordering. We think these results on condition number are acceptable—at least, for every problem the majority of the algorithms and preorderings produced well-conditioned bases—but we do not know how to guarantee good conditioning. Coleman and Pothen [3] gave an algorithm for finding an orthogonal basis for the null space, but they also showed that the sparsest orthogonal null basis may be very much denser than an arbitrary null basis. How to trade off sparsity for conditioning is an interesting open question.

Coleman and Pothen [4] are experimenting with the null basis algorithms from Pothen's thesis, but we do not yet have any comparisons with our algorithms. Pothen [14] has also recently suggested an interesting class of heuristics for null basis problems from structural analysis, based on the structure of the object being analyzed in addition to the structure of the matrix.

Appendix. Bipartite graphs, matchings and rank. Let A be a matrix. The *bipartite graph of* A is the graph whose vertices are the rows of A and the columns of A, with an edge between a row vertex and a column vertex if and only if the corresponding entry of A is nonzero. We do not distinguish between a row of A and a row vertex of the graph of A. Informally, we do not distinguish between A and its graph.

A *matching* on A is a set of edges, no two of which have a common endpoint. (Equivalently, it is a set of nonzeros, no two of which are in the same row or column.) A vertex is *covered* by a matching if it is the endpoint of some matching edge, and *uncovered* otherwise. A *maximum matching* is a matching such that no matching on A has more edges.

A *path* in A is a sequence of distinct vertices $v_0, v_1, \cdots, v_k$ such that $\{v_{i-1}, v_1\}$ is an edge for $1 \leqq i \leqq k$. The *length* of the path is k. If M is a matching on A, an *alternating path* (with respect to M) is a path whose edges are alternately in M and not in M. If $P = v_0, \cdots, v_k$ is an alternating path from an uncovered vertex v_0 to an uncovered vertex v_k, we can modify the matching by removing edges $\{v_1, v_2\}, \{v_3, v_4\}, \cdots, \{v_{k-2}, v_{k-1}\}$ and adding edges $\{v_0, v_1\}, \{v_2, v_3\}, \cdots, \{v_{k-1}, v_k\}$. This is called *alternating along the path P*. It increases the size of the matching by one edge, covers v_0 and v_k, and leaves all previously covered vertices covered.

For proofs of Lemmas 1 and 2 below and more background on bipartite matching, see Lawler [12] or Papadimitriou and Steiglitz [13].

LEMMA 1 (Hall's Theorem). *Matrix A has a matching that covers every column if and only if every set of columns of A intersects a set of rows of A that is at least as large.*

LEMMA 2. *If M is a matching on A whose size is not maximum, then there is an alternating path from some uncovered row of A to some uncovered column of A.*

LEMMA 3. *Every matrix has a matching that is at least as large as its numerical rank.*

Proof. Let k be the numerical rank of matrix A. Let B be a submatrix consisting of k linearly independent columns of A. Any set of $t \leqq k$ columns of B is a submatrix of full column rank, so it must have nonzeros in at least t rows. By Hall's theorem, B has a matching that covers all k columns, so A has a matching of size at least k. □

The combinatorial notion of maximum matching size corresponds closely to the numerical notion of rank. It can be shown that if we fix the nonzero structure of A and assign values to those nonzeros at random, then with probability 1 the rank is equal to the maximum matching size.

REFERENCES

[1] M. BERRY, M. HEATH, I. KANEKO, M. LAWO, R. PLEMMONS AND R. WARD, *An algorithm to compute a sparse basis of the null space*, Numer. Math., 47 (1985), pp. 483–504.

[2] T. F. COLEMAN, *Large Sparse Numerical Optimization*, Springer-Verlag, New York, 1984.

[3] T. F. COLEMAN AND A. POTHEN, *The sparse null space basis problem* I: *Complexity*, this Journal, 7 (1986), pp. 527–537.

[4] ———, *The sparse null space basis problem* II: *Algorithms*, Rpt. CS-86-747, Cornell University, Ithaca, NY, 1986.

[5] I. S. DUFF, *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7 (1981), pp. 315–330.

[6] A. GEORGE AND J. W. H. LIU, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[7] P. E. GILL, W. MURRAY AND M. H. WRIGHT, *Practical Optimization*, Academic Press, New York, 1981.

[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[9] M. T. HEATH, *Some extensions of an algorithm for sparse linear least squares problems*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 223–237.

[10] M. T. HEATH, R. J. PLEMMONS AND R. C. WARD, *Sparse orthogonal schemes for structural optimization using the force method*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 514–532.

[11] I. KANEKO, M. LAWO AND G. THIERAUF, *On computational procedures for the force method*, Internat. J. Numer. Meth. Engrg., 18 (1982), pp. 1469–1495.

[12] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.

[13] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[14] A. POTHEN, *Equilibrium graphs in structural optimization*, Rpt. CS-86-22, Pennsylvania State University, University Park, PA, 1986.

[15] A. TOPCU, *A contribution to the systematic analysis of finite element structures using the force method*, Ph.D. thesis, University of Essen, Germany, 1979. (In German.)

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

A CLASS OF EFFICIENT VALUES FOR GAMES IN PARTITION FUNCTION FORM*

EDWARD M. BOLGER†

Abstract. In this paper we derive a class of linear, efficient, dummy-independent values for n-person games in partition function form. Each of these values is an extension of the Shapley value.

Key words. game theory, value theory

AMS(MOS) subject classification. 90

1. Introduction. Games in partition function form were introduced in Lucas and Thrall [2] to generalize games in characteristic function form. Myerson [3] derived an efficient value for games in partition function form which is a natural extension of the Shapley [4] value.

Games in partition function form can be used to model r-candidate voting games. In Bolger [1], an efficient linear index which assigns value 0 to dummy players is derived for such games. This index can be used as a value for games in partition function form. In this paper we derive a class of efficient linear values for games in partition function form.

DEFINITION 1. Let $N = \{1, 2, 3, \cdots, n\}$ and let Γ be a partition of N into nonempty subsets of N. Let $S \in \Gamma$. We call (S, Γ) an *embedded coalition* (ECL).

DEFINITION 2. A real-valued function v whose domain is the set of all ECL's is called a *characteristic function*. The pair (N, v) is called a *game on N in partition function form*. By convention, we set $v(\phi, \Gamma) = 0$ for each game v and each partition Γ.

The collection of all such games is denoted by R^{ECL}. Considering R^{ECL} as a vector space over R, Myerson [3] observed that its dimension equals the number of ECL's.

Let (S, Γ) be an ECL. Define the game $v^{S,\Gamma}$ by

$$v^{S,\Gamma}(S', \Gamma') = \begin{cases} 1 & \text{if } (S', \Gamma') = (S, \Gamma), \\ 0 & \text{otherwise.} \end{cases}$$

As a basis for R^{ECL}, we can use the collection $\{v^{S,\Gamma} : (S, \Gamma) \text{ is an ECL}\}$.

For the ECL (S, Γ), $|S|$ denotes the cardinality of S and $|\Gamma|$ denotes the cardinality of Γ.

DEFINITION 3. Let π be a permutation of N. The game πv is defined by

$$\pi v(S, \Gamma) = v(\pi S, \pi \Gamma).$$

2. Main results. The Shapley value, Φ, for games in characteristic function form can be written in the form:

$$\Phi_i(v) = \sum_{\substack{S \subset N \\ i \in S}} \frac{(|S|-1)!(n-|S|)!}{n!} v(S) - \sum_{\substack{S \subset N \\ i \notin S}} \frac{(|S|)!(n-|S|-1)!}{n!} v(S).$$

This Shapley value could be applied to games in partition function form; that is,

* Received by the editors May 27, 1986; accepted for publication December 24, 1986.

† Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056.

one could use the value

$$\theta_i(v) = \sum_{\substack{(S,\Gamma)\\ i\in S}} \frac{(|S|-1)!(n-|S|)!}{n!} v(S,\Gamma) - \sum_{\substack{(S,\Gamma)\\ i\notin S}} \frac{(|S|)!(n-|S|-1)!}{n!} v(S,\Gamma).$$

Theorem 1 will show that this value is efficient. However, it sometimes assigns nonzero values to dummy players and zero value to nondummy players in monotonic simple games. Moreover, for each $i \in N$, $\theta_i(v^{S,\Gamma}) = \theta_i(v^{S,\Gamma'})$ whenever $S \in \Gamma \cap \Gamma'$.

In our search for a value θ, we shall relax this last property slightly by requiring instead that for i in N, the value θ satisfy

Axiom 1. $\theta_i(v^{S,\Gamma}) = \theta_i(v^{S,\Gamma'})$ whenever $S \in \Gamma \cap \Gamma'$ and $|\Gamma| = |\Gamma'|$.

The remaining five axioms will be the familiar symmetry, linearity, efficiency and dummy axioms.

Axiom 2. $\theta_{\pi i}(v) = \theta_i(\pi v)$.

Axiom 3. $\theta(v + w) = \theta(v) + \theta(w)$.

Axiom 4. $\theta_1(v) + \cdots + \theta_n(v) = v(N, \{N\})$.

We will begin with some results based on the first four axioms.

THEOREM 1. *Let Φ be a value on the class of games on N in partition function form. Then Φ satisfies Axioms* 1–4 *if and only if there is a function $a(|S|, n, |\Gamma|)$ such that*

$$\Phi_i(v) = \sum_{\substack{(S,\Gamma)\\ i\in S}} a(|S|,n,|\Gamma|)v(S,\Gamma) - \sum_{\substack{(S,\Gamma)\\ i\notin S}} \frac{|S|}{n-|S|} a(|S|,n,|\Gamma|)v(S,\Gamma) \tag{1}$$

and

$$a(n,n,1) = \frac{1}{n}. \tag{2}$$

Proof. We shall prove the necessity and leave the proof of the sufficiency for the reader. For fixed t and m, let $S' = \{1, 2, \cdots, t\}$ and

$$\Gamma' = \{\{1,2,\cdots,t\},\{t+1,\cdots,n-m+2\},\{n-m+3\},\cdots\{n\}\}.$$

Now define

$$a(t,n,m) = \Phi_1(v^{S',\Gamma'}) \quad \text{and} \quad b(t,n,m) = -\Phi_n(v^{S',\Gamma'})$$

and let

$$\theta_i(v) = \sum_{\substack{(S,\Gamma)\\ i\in S}} a(|S|,n,|\Gamma|)v(S,\Gamma) - \sum_{\substack{(S,\Gamma)\\ i\notin S}} b(|S|,n,|\Gamma|)v(S,\Gamma).$$

Clearly $\theta(v^{S',\Gamma'}) = \Phi(v^{S',\Gamma'})$.

Now let (S, Γ) be an ECL. We can write $\Gamma = \{S, T_1, \cdots, T_k, \{j\}, \cdots, \{n\}\}$ where $|T_i| > 1$ for each i and where $j = n - |\Gamma| + k + 2$. Form a new partition Γ_1 by removing all but one player from each of $T_2, \cdots, T_k$ and placing these removed players in with the members of T_1. Since $|\Gamma| = |\Gamma_1|$, Axiom 1 guarantees that $\Phi(v^{S,\Gamma}) = \Phi(v^{S,\Gamma_1}) = \theta(v^{S,\Gamma_1}) = \theta(v^{S,\Gamma})$. Since Φ and θ agree on $\{v^{S,\Gamma} : (S, \Gamma) \in \text{ECL}\}$, Φ and θ agree for all games on N in partition function form by Axiom 3.

Finally, if we apply Axiom 4 to $v^{S,\Gamma}$, we get immediately that for $|S| < n$,

$$b(|S|,n,|\Gamma|) = \frac{|S|}{n-|S|} a(|S|,n,|\Gamma|).$$

In the sequel, we shall only consider values satisfying Axioms 1–4.

DEFINITION 4. Player j is called a dummy in the game v if for each nontrivial ECL (S, Γ), $v(S, \Gamma) = v(S - \{j\}, \Gamma')$ for every Γ' where Γ' is a partition of N obtained from Γ by moving player j to some other set in Γ or to a new set by itself.

It is clear from Theorem 1 that there are many efficient, linear, symmetric values for games in partition function form. We shall now impose the restriction that θ shall assign value 0 to each dummy.

Axiom 5. If player j is a dummy, then $\theta_j(v) = 0$.

DEFINITION 5. For a fixed (S, Γ), let $v = v^{S,\Gamma}$ and for $d \notin N$, define the dummy extension v^d on $N \cup \{d\}$ by the following:

(i) $v^d(S, \Gamma^d) = 1$ for each partition Γ^d of $N \cup \{d\}$ obtained by placing d in some set of Γ other than S or in a set by itself.

(ii) $v^d(S \cup \{d\}, (\Gamma - \{S\}) \cup (S \cup \{d\})) = 1$.

(iii) $v^d(S', \Gamma') = 0$ for all other (S', Γ') where Γ' is a partition of $N \cup \{d\}$.

An easy calculation yields

$$\theta_d(v^d) = a(|S|+1, n+1, |\Gamma|) - (|\Gamma|-1)b(|S|, n+1, |\Gamma|) - b(|S|, n+1, |\Gamma|+1).$$

We get immediately the following theorem.

THEOREM 2. *Φ satisfies Axioms* 1–5, *if and only if there is a function $a(|S|, n, |\Gamma|)$ satisfying* (1) *and* (2) *above and*

$$\begin{aligned} a(|S|+1, n, |\Gamma|) &= \frac{(|\Gamma|-1)|S|}{(n-|S|)} a(|S|, n, |\Gamma|) \\ &+ \frac{|S|}{n-|S|} a(|S|, n, |\Gamma|+1) \quad \textit{for } |\Gamma| \leqq n-1. \end{aligned} \tag{3}$$

COROLLARY 1. *If Φ satisfies Axioms* 1–5, *then*

$$a(n-1, n, 2) = [n(n-1)]^{-1}.$$

Proof. Put $|S| = n - 1$ in Theorem 2.

Example. The function

$$a(t, n, m) = \frac{(t-1)!(n-t)!(n-1)!}{n!(n-1)^{n-t}(n-m)!}$$

determines a value satisfying Axioms 1–5.

We proceed now to show that Axioms 1–5 determine an $n - 2$ parameter family of values. First we prove two combinatorial lemmas.

LEMMA 1. *If $j \leqq q - 1$, then*

$$\sum_{l=1}^{q+1} \frac{(-1)^{l-1} l^j}{(l-1)!(q+1-l)!} = 0.$$

Proof. The result is true for $j = 0$ since $(1 + (-1))^q = 0$. Assume that the result is true up to $j - 1$. Then

$$\begin{aligned} \sum_{l=1}^{q+1} \frac{(-1)^{l-1} l^j}{(l-1)!(q+1-l)!} &= \sum_{i=0}^{q} \frac{(-1)^i (1+i)^j}{i!(q-i)!} \\ &= \sum_{i=0}^{q} \frac{(-1)^i}{i!(q-i)!} \sum_{k=0}^{j} \binom{j}{k} i^k \\ &= \sum_{k=0}^{j} \binom{j}{k} \sum_{i=1}^{q} \frac{(-1)^i i^{k-1}}{(i-1)!(q-i)!} = 0 \end{aligned}$$

by the induction hypothesis since $k - 1 \leqq j - 1 \leqq q - 2$.

LEMMA 2. *If* $0 \leqq m \leqq q-1$, *then*

$$\sum_{l=1}^{q}(-1)^{l-1}\frac{(n-l+q)^m}{(l-1)!(q+1-l)!}=(-1)^{q+1}\frac{(n-1)^m}{q!}.$$

Proof.

$$\sum_{l=1}^{q+1}(-1)^{l-1}\frac{(n-l+q)^m}{(l-1)!(q+1-l)!}=\sum_{l=1}^{q+1}(-1)^{l-1}\sum_{j=0}^{m}\binom{m}{j}\frac{(n+q)^{m-j}(-l)^j}{(l-1)!(q+1-l)!}$$

$$=\sum_{j=0}^{m}\binom{m}{j}(n+q)^{m-j}(-1)^j\sum_{l=1}^{q+1}\frac{(-1)^{l-1}l^j}{(l-1)!(q+1-l)!}$$

$$=0 \text{ by Lemma 1.}$$

Now for positive integers m and i, let

$$P(m,i)=m(m-1)\cdots(m-i+1) \quad \text{for } i \leqq m,$$

$$U(m,i)=m(m+1)\cdots(m+i-1),$$

$$P(m,0)=U(m,0)=1.$$

Furthermore, if i is a negative integer, we set

$$\frac{P(m,i)}{U(m,i)}=0$$

and interpret all sums of the form $\sum_{q=0}^{i}$ to be 0. These conventions will also enable us to avoid clumsy special cases in the succeeding results.

THEOREM 3. *Equation* (3) *holds if and only if for* $-1 \leqq j < k$,

$$a(k-j,n,n-k)=\sum_{p=0}^{k}\sum_{l=1}^{p+1}\frac{P(n-k+j,j-p+1)}{U(k-j,j-p+1)} \cdot\frac{(n-k+p-1)}{(n-k-l+p)^{j+2-p}}\frac{a(1+k-p,n,n-k+p)}{(-1)^{l-1}(l-1)!(p+1-l)!}. \tag{4}$$

Proof. By (3), Theorem 3 is true for $k=1$. Assume inductively that the theorem is true up to and including $k-1$.

For simplicity of notation, let

$$F(q,l)=(-1)^{l-1}(l-1)!(q+1-l)!$$

and

$$A(t,m)=a(t,n,m).$$

By successive application of Theorem 2,

$$A(k-j,n-k)=\frac{P(n-k+j,j+1)A(k+1,n-k)}{(n-k-1)^{j+1}U(k-j,j+1)}$$

$$-\sum_{i=0}^{j}\frac{P(n-k+j,j-i)A(k-i,n-k+1)}{(n-k-1)^{j+1-i}U(k-j,j-i)}$$

$$=\frac{P(n-k+j,j+1)A(k+1,n-k)}{(n-k-1)^{j+1}U(k-j,j+1)}$$

$$- \sum_{i=0}^{j} \frac{P(n-k+j,j-i)}{(n-k-1)^{j+1-i}U(k-j,j-i)}$$

$$\cdot \sum_{p=0}^{k-1} \sum_{l=1}^{p+1} \frac{P(n-k+i,i-p)}{U(k-i,i-p)} \frac{(n-k+p)A(k-p,n-k+1+p)}{(n-k+1-l+p)^{i-p+1}F(p,l)}$$

$$= \frac{P(n-k+j,j+1)A(k+1,n-k)}{(n-k-1)^{j+1}U(k-j,j+1)}$$

$$- \sum_{p=0}^{k-1} \sum_{l=1}^{p+1} \frac{P(n-k+j,j-p)}{U(k-j,j-p)} \frac{(n-k+p)A(k-p,n-k+1+p)}{(n-k+1-l+p)^{-p}F(p,l)}$$

$$\cdot \sum_{i=0}^{j} \frac{(n-k+1-l+p)^{-(i+1)}}{(n-k-1)^{j+1-i}}$$

$$= \frac{P(n-k+j,j+1)A(k+1,n-k)}{(n-k-1)^{j+1}U(k-j,j+1)}$$

$$- \sum_{p=0}^{k-1} \sum_{l=1}^{p+1} \frac{P(n-k+j,j-p)}{U(k-j,j-p)} \frac{(n-k+p)A(k-p,n-k+1+p)}{(n-k+1-l+p)^{-p}F(p,l)(l-p-2)}$$

$$\cdot \left[\left(\frac{1}{n-k+1-l+p} \right)^{j+1} - \left(\frac{1}{n-k-1} \right)^{j+1} \right]$$

$$= \frac{P(n-k+j,j+1)A(k+1,n-k)}{(n-k-1)^{j+1}U(k-j,j+1)}$$

$$+ \sum_{q=1}^{k} \sum_{l=1}^{q} \frac{P(n-k+j,j-q+1)}{U(k-j,j-q+1)} \frac{(n-k+q-1)A(1+k-q,n-k+q)}{(n-k-l+q)^{j+2-q}F(q,l)}$$

$$- \frac{1}{(n-k-1)^{j+1}} \sum_{q=1}^{k} \sum_{l=1}^{q} \frac{P(n-k+j,j-q+1)}{U(k-j,j-q+1)}$$

$$\cdot \frac{(n-k+q-1)A(1+k-q,n-k+q)}{(n-k-l+q)^{1-q}F(q,l)}$$

$$= \frac{P(n-k+j,j+1)A(k+1,n-k)}{(n-k-1)^{j+1}U(k-j,j+1)}$$

$$+ \sum_{q=1}^{k} \sum_{l=1}^{q+1} \frac{P(n-k+j,j-q+1)}{U(k-j,j-q+1)} \frac{(n-k+q-1)A(1+k-q,n-k+q)}{(n-k-l+q)^{j+2-q}F(q,l)}$$

by Lemma 2. This completes the proof of the necessity. The proof of the sufficiency is left to the reader.

Remarks. Suppose $n = 4$. Then the possible values of $4 - (k - p)$ will be 4, 3, 2, 1. If θ satisfies Axioms 1-5, then since the values of $a(4, 4, 1)$ and $a(3, 4, 2)$ are known (Theorem 1 and Corollary 1), we need only specify the values of $a(2, 4, 3)$ and $a(1, 4, 4)$ to determine all relevant values of the function a. That is, for $n = 4$, we get a two-parameter family of power indices θ which are efficient and satisfy the dummy axiom. Specifically, if we let $x = a(2, 4, 3)$ and $y = a(1, 4, 4)$, then $a(2, 4, 2) = -x + \frac{1}{12}$, $a(1, 4, 3) = 1.5x - 0.5y$, $a(1, 4, 2) = 0.25 - 4.5x + 0.5y$.

Example. Let $n = 3$ and let v be the game on N defined by $v(N, \{N\}) = v(\{1, 2\}, \{\{1, 2\}, \{3\}\}) = v(\{1, 3\}, \{\{1, 3\}, \{2\}\}) = v(\{1\}, \{\{1\}, \{2\}, \{3\}\}) = 1$; $v(S, \Gamma) = 0$ for all other (S, Γ). If we choose $a(1, 3, 3) = 0.10$, then $\theta_1(v) = .7667$. Suppose we add a dummy player 4 to the game v. Then if we let v^4 be the dummy extension of v, $\theta_1(v^4)$ will also equal 0.7667 provided we choose $a(2, 4, 3)$ to equal 0.025.

The above example shows that the value of the other players may be affected by adding a dummy player to the game. We seek now conditions on the function a which will make θ "dummy-independent."

For each game v on N and $d \notin N$, we have the following axiom.

Axiom 6. $\theta_i(v^d) = \theta_i(v)$ for $i \in N$.

LEMMA 3. *If θ satisfies Axioms* 1–6, *then the function $a(|S|, n, |\Gamma|)$ in* (1) *satisfies*

$$(5)\quad a(|S|, n, |\Gamma|) = (|\Gamma| - 1)a(|S|, n+1, |\Gamma|) + a(|S|+1, n+1, |\Gamma|) + a(|S|, n+1, |\Gamma|+1),$$

$$(6)\quad b(|S|, n, |\Gamma|) = (|\Gamma| - 1)b(|S|, n+1, |\Gamma|) + b(|S|+1, n+1, |\Gamma|) + b(|S|, n+1, |\Gamma|+1)$$

where $b(|S|, n, |\Gamma|) = |S|a(|S|, n, |\Gamma|)/(n - |S|)$.

Proof. For a fixed (S, Γ), let $v = v^{S,\Gamma}$ and let v^d be the dummy-extension of v. Write $|\Gamma| = m$. For $i \in S$,

$$\theta_i(v^d) = (m-1)a(|S|, n+1, m) + a(|S|+1, n+1, m) + a(|S|, n+1, m+1)$$

whereas for $i \notin S$, $i \neq d$,

$$\theta_i(v^d) = -(m-1)b(|S|, n+1, m) - b(|S|+1, n+1, m) - b(|S|, n+1, m+1).$$

LEMMA 4. *If θ satisfies Axioms* 1–6, *then*

$$a(n - |\Gamma| + 2, n+1, |\Gamma|) = \frac{n - |\Gamma| + 1}{n+1} a(n - |\Gamma| + 1, n, |\Gamma|).$$

Proof. Write $|\Gamma| = m$. From Theorem 2,

$$a(|S|, n+1, m) = \frac{(n+1-|S|)a(|S|+1, n+1, m)}{(m-1)|S|} - \frac{1}{m-1} a(|S|, n+1, m+1).$$

Using this in (6) we get, since $b(t, n, m) = ta(t, n, m)/(n - t)$,

$$a(|S|, n, m) = \frac{n+1}{|S|} a(|S|+1, n+1, m).$$

Repeated applications of Lemma 4 yield the following theorem.

THEOREM 4. *θ satisfies Axioms* 1–6 *if and only if there is a function $a(|S|, n, |\Gamma|)$ satisfying* (1), (2), (4) *and*

$$a(n - m + 1, n, m) = \frac{(n-m)!}{P(n, n-m)} a(1, m, m).$$

Consequently, if θ satisfies Axioms 1–6, then the values of θ are determined by the choice of $a(1, n, n)$, $a(1, n-1, n-1)$, $\cdots$, $a(1, 3, 3)$. In particular, for $n = 4$, the choice of $a(1, 4, 4)$ and $a(1, 3, 3)$ determines the value for all two, three and four player games.

3. A final note. For fixed $r > 0$, consider the collection of games in partition function form for which $v(S, \Gamma) = 0$ if $|\Gamma| > r$. This collection of games may serve as models for voting games in which a set of voters will select one of r candidates. These voting games are determined by specifying for each ECL (S, Γ) whether or not S wins with respect to Γ.

For the above collection of games, a value satisfying Axioms 1–6 is determined by

$$a(t, n, m) = \frac{(t-1)!(n-t)!(r-1)!}{n!(r-1)^{n-t}(r-m)!}.$$

REFERENCES

[1] E. BOLGER, *Power indices for multicandidate voting games*, Internat. J. Game Theory, 15 (1986), pp. 175–186.

[2] W. F. LUCAS AND R. M. THRALL, *n-person games in partition function form*, Naval Res. Logist. Quart., X (1963), pp. 281–298.

[3] R. B. MYERSON, *Values of games in partition function form*, Internat. J. Game Theory, 6 (1977), pp. 23–31.

[4] L. S. SHAPLEY, *A value for n-person games*, in Contributions to the Theory of Games II, H. W. Kuhn and A. W. Tucker, eds., Princeton Univ. Press, Princeton, NJ, 1953, pp. 307–317.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

014

THE STRUCTURE OF MONOMIAL CIRCULANT MATRICES*

WILLIAM C. WATERHOUSE†

Abstract. Consider the generalized circulant matrices recently introduced by P. J. Davis and K. Wang, where row i is obtained from row 1 by permuting the entries and also multiplying by scaling factors. When such families are (like circulants) closed under multiplication, they can be reduced to certain standard forms that are in fact related to twisted group algebras and group cohomology. For real matrices or complex matrices, there are only finitely many such standard forms for each size. Similar results hold for the appropriate generalizations of g-circulants. In particular, in each case (for complex matrices) we can describe the determinant as a function of the first-row entries: it is a product of powers of various smaller determinants, and up to (explicit) scaling factors the list of entries in these determinants is a unitary transformation of the original first-row entries. Some of the ideas hold for even more generalized forms of circulants.

Key words. circulant matrices, generalized circulants, monomial matrices, twisted group algebras, group cohomology, algebras of matrices

AMS(MOS) subject classifications. 15A30, 16A46, 20C25, 20J06

The varied uses discovered for circulant matrices, ranging from physics to combinatorics with intermediate stops at geometry and statistics (cf. [6]), amply justify generalization—provided that the generalizations have a similarly reasonable and coherent structure. Group matrices and g-circulants, for instance, are reasonable and (correspondingly) useful; matrices constructed by random permutations of the first row have nothing to recommend them. One helpful guideline has been to consider behavior under matrix multiplication: group matrices, like circulants, are closed under multiplication, and g-circulants are at least sent to other g-circulants when multiplied by ordinary circulants. It is in fact true [14] that a suitable combination of these two types gives all possible families of matrices that are constructed by permuting the first row and are well behaved under multiplication.

Just recently, Wang and Davis [13] introduced a far-reaching generalization of circulants where the entries are derived from the first row by permuting the elements *and* multiplying by fixed (nonzero) constants depending on the position. Another way of saying this is that row i comes from row 1 under multiplication by a monomial matrix (one with just one nonzero entry in each row and column), and hence I propose to call such families of matrices **monomial circulants.** The goal of this paper is to analyze and classify those families of monomial circulants that are well behaved under multiplication. Wang and Davis began such an analysis, showing that the families that formed algebras could all be constructed as the "λ-group matrices" introduced earlier by Wang [11]. But the definition of λ-group matrices allows the same family of matrices to arise from quite different input data, and thus their results do not make clear how many essentially different types of such circulants exist.

It turns out that the correct analysis uses what are called "twisted group algebras." Fortunately, this paper can be understood with no previous knowledge of that topic. Readers who want a quick idea of the types of matrices involved over the real numbers can turn immediately to § 4, where the 4 by 4 families are written out in standard form. The analysis itself begins in § 1 by showing how the λ-group matrices of [13] lead us to twisted group algebras. Section 2 shows more generally how monomial circulant families

* Received by the editors June 18, 1986; accepted for publication December 24, 1986. This work was supported in part by the National Science Foundation under grant DMS 8400649.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

closed under multiplication (but perhaps not including the identity) can be reduced to twisted group algebras. Section 3 lists some of the results that follow from this analysis, information not at all apparent from the original definition of the circulants. We shall see, for instance, how the determinant depends on the entries in the first row. Furthermore, over the real or complex numbers there will be only finitely many equivalence classes of any given size. Section 5 contains a selfadjointness result yielding an improved block decomposition theorem (which Wang and Davis proved in the commutative case). The final sections characterize the analogue of g-circulants associated with monomial circulant algebras and show how their equivalence classes can be computed in terms of algebra automorphisms.

To emphasize the underlying algebra, I have included an Appendix briefly discussing the "ultimate" generalization of circulants, where row i is derived from the first row u as $uC[i]$ for some arbitrary (fixed) set of invertible matrices $C[i]$; we might call these **linear circulants.** Even the families of this general type closed under multiplication turn out to be objects already studied in ring theory (and they have a surprisingly large amount of structure; see [9, pp. 445–455] or [5]).

1. The meaning of λ-group matrices. To make contact with the earlier work, we begin by analyzing the definition of λ-group matrices, as given in [13]. The definition starts with a finite group G of order n and a central group extension E,

$$1 \to H \to E \to G \to 1.$$

There is also a specified homomorphism λ from H to the multiplicative group of some base field k. The "λ-group matrices" over k are then the matrices of the form

$$A = \phi(t_i^{-1} t_j)$$

where $t_1 = 1, t_2, \cdots, t_n$ are coset representatives for H in E and ϕ runs over all "λ-maps," maps $E \to k$ satisfying the condition $\phi(he) = \lambda(h)\phi(e)$ for all h in H and e in E. (There is a more special but equivalent definition of λ-group matrices in [11] for the case when H is finite.) Our goal in this section is to show how these matrices can be put in a more recognizable form.

For convenience, we identify G with the set of indices of the t_i (so 1 becomes the identity of G); this identification can best be interpreted as introducing a group structure on the set of indices. Now we choose a nice basis ϕ_g for the λ-maps $E \to k$: clearly any λ-map ϕ is determined by the values $\phi(t_i)$, and we let $\phi_g(t_i)$ be 1 when $i = g$ and 0 otherwise. When we compute the matrix $A[g]$ corresponding to ϕ_g, we find that

$$A[g]_{ij} = \delta_{ig,j}\lambda(h(i,g)).$$

By construction these matrices are a basis for the λ-group matrices. An arbitrary λ-group matrix with first row entry (say) $c[g]$ in column g is equal to $\Sigma c[g]A[g]$, and its entry in some subsequent row i and column j is given by

$$\Sigma c[g]A[g]_{ij} = \lambda(h(i, i^{-1}j))c[i^{-1}j].$$

Thus we do indeed have monomial circulants.

Associativity in the group E tells us that $t_i(t_g t_p) = (t_i t_g)t_p$, so if we write $\beta(i, g) = \lambda(h(i, g))$, we have the identity

$$(*) \qquad \beta(i,g)\beta(ig,p) = \beta(i,gp)\beta(g,p).$$

This says that β is a "two-cocycle" of G with values in the multiplicative group k^*. (It is a "normalized" cocycle, which is to say that $\beta(1, p) = \beta(p, 1) = 1$ for all p in G; this just corresponds to the choice of $t_1 = 1$ in E.) Now using (*) we can compute that our matrices satisfy the basic multiplication rule

$$(**) \qquad A[g]A[p] = \beta(g,p)A[gp].$$

Some readers may recall that similar formulas occur in the (related) topic of projective representations.

The algebra of dimension n over k with one basis element for each g in G and multiplication given by the "twisted" group multiplication (**) is what is called the twisted group algebra for the cocycle β, denoted $k[G, \beta]$. Thus we see that the algebra of λ-group matrices is isomorphic to the twisted group algebra for the cocycle $\beta(g, p) = \lambda(t_{gp}^{-1} t_g t_p)$. But actually we have a more explicit expression. If e_s is the row from k^n with a single nonzero entry equal to 1 in column s, the formula for $A[g]$ gives us $e_sA[g] = \beta(s, g)e_{sg}$. Comparing this with the formula (**) for right multiplication in the algebra, we have our first theorem.

THEOREM 1. *A family of λ-group matrices consists precisely of the matrices expressing a right regular representation of a twisted group algebra in its natural (group-element) basis.* □

Note that the original group extension E was scaffolding that has now been removed from the finished construction: the λ-group matrices depend only on the normalized cocycle β that we derived from the original data. Furthermore, any normalized cocycle can occur. Indeed, it occurs with λ = identity and $H = k^*$, as follows from the standard theory of group extensions [5]. The role of the homomorphism $\lambda : H \to k^*$ was merely to move a cocycle with values in H over to the cocycle β with values in k^*. Thus the possible families of λ-group matrices for a given index group G correspond precisely to the normalized 2-cocycles of G with values in the multiplicative group k^*.

There is a noncanonical choice involved in the definition of λ-group matrices: they depend not only on G and the extension but also on the choice of coset representatives t_g. If we replace t_g by $t_gh[g]$ for some (arbitrary) elements $h[g]$ in H, then we have

$$t_{gp}h[gp] = \{h[gp]h[g]^{-1}h[p]^{-1}\}h(g,p)(t_gh[g])(t_ph[p]).$$

Thus if we let $\gamma(g) = \lambda(h[g])$ in k^*, we will replace $\beta(g, p)$ by

$$\beta'(g,p) = \gamma(gp)\gamma(g)^{-1}\gamma(p)^{-1}\beta(g,p),$$

which in the cohomology theory is called changing β to a cohomologous cocycle. With an appropriately chosen extension we can change β to any such cocycle. (To keep it normalized, we just need to have $\gamma(1) = 1$, which corresponds to maintaining the original condition $t_1 = 1$.) Then in the family of matrices for β', the matrix $\Sigma c'[g]A'[g]$ with first row $(c'[g])$ will have its (i, j)-entry equal to

$$\beta'(i, i^{-1}j)c'[i^{-1}j] = \gamma(j)\gamma(i)^{-1}\gamma(i^{-1}j)^{-1}\beta(i, i^{-1}j)c'[i^{-1}j].$$

Thus if we write $c[g] = c'[g]\gamma(g)^{-1}$ and let D be the diagonal matrix with entries $\gamma[g]^{-1}$, we have

$$\Sigma c'[g]A'[g] = D(\Sigma c[g]A[g])D^{-1}.$$

That is, one family is taken to the other by conjugation by a diagonal matrix. Clearly the computation can be reversed. Thus we have a further correspondence.

THEOREM 2. *Two families of λ-group matrices for the same index group G but different cocycles differ by a diagonal conjugation iff the cocycles are cohomologous.* □

2. Algebras of monomial circulants. Wang and Davis showed [13] that any family of monomial circulants which is an algebra (i.e., contains the identity and is closed under multiplication) is actually a family of λ-group matrices. In view of the results of § 1, this theorem is included in the following result.

THEOREM 3. *Suppose a family of monomial circulants is closed under multiplication. Then there is a composition law on the index set making it a semigroup with bijective left multiplications, and the family is the right regular representation for a twisted semigroup algebra on this semigroup. If the identity matrix is in the family, then we have a twisted group algebra.*

Proof. Take the element $A[p]$ in the family whose only nonzero entry in row 1 is a 1 in column p. The monomial property says that in row i it has an entry in just one column. We denote the column involved by ip and the entry by $\beta(i, p)$. Thus in the law of formation for these monomial circulants, $p \mapsto ip$ is the permutation involved in row i, and the $\beta(i, p)$ are the multipliers. By definition then $1p = p$ for all p, the operation $p \mapsto ip$ for fixed i is bijective, and $\beta(1, p) = 1$. The product $A[p]A[q]$ has only one nonzero entry in row 1, namely $\beta(p, q)$ in column pq. Since this product must be in the family, we get $A[p]A[q] = \beta(p, q)A[pq]$. Comparing the entries in row i on each side of this equation, we first find (by seeing where they are nonzero) that $(ip)q = i(pq)$; this shows we have a semigroup of the type required. Then by comparing the nonzero entries, we find that

$$\beta(i, pq)\beta(p, q) = \beta(i, p)\beta(ip, q),$$

precisely the cocycle identity that yields associativity for the twisted semigroup algebra. If the identity matrix is in the family, it must be given by $A[1]$, and hence also $i1 = i$ for all i; thus since 1 is a two-sided identity and $p \mapsto ip$ is bijective, the index set is a group. (In this case we also have $\beta(i, 1) = 1$.) □

Clearly we should say that two families of monomial circulants are (monomially) *equivalent* if one can be taken to the other under conjugation by a monomial matrix. Using that equivalence relation, we can in essence reduce the general case of Theorem 3 to twisted group algebras.

THEOREM 4. *Take any family of monomial circulants closed under multiplication. Then (up to monomial equivalence) the matrices in it have the form*

$$\begin{bmatrix} M_1 M_2 \cdots M_s \\ M_1 M_2 \cdots M_s \\ \cdots\cdots\cdots \\ M_1 M_2 \cdots M_s \end{bmatrix}$$

where all the M_i are monomial circulants coming from the same twisted group algebra.

Proof. The structure of semigroups of the type occurring here is known [8, p. 54] (the proof is also reproduced in [14]). The subset G of indices for which $g1 = g$ is a group. There is also a subset S of elements s for which $sp = p$ for all p, and every element in the semigroup is uniquely expressed as a product gs. We shall build our block decomposition on this, letting each block consist of row and column indices with fixed S-factors. We also arrange the entries in the different blocks with their G-factors always in the same order. This all just permutes the original index set, and hence it changes the family only up to equivalence.

The cocycle identity for g in G gives us

$$\beta(g^{-1}, g)\beta(g^{-1}g, 1) = \beta(g^{-1}, g1)\beta(g, 1),$$

which yields $\beta(g, 1) = 1$ for all such g. Define now in general $\gamma(gs)$ to be $\beta(gs, 1)$. Since $\gamma(1) = \gamma(g) = 1$, we have $\gamma(gs)\gamma(1)\gamma(gs1)^{-1} = \beta(gs, 1)$. By Theorem 2, a cohomology change of β corresponds to a diagonal conjugation, so after such a conjugation we can assume that $\beta(gs, 1) = 1$ for all gs.

It follows now that $\beta(gs, t) = \beta(gs, t)\beta(gst, 1) = \beta(gs, t1)\beta(t, 1) = 1$ for all t in S. From this it follows that $\beta(ht, gs) = \beta(ht, gs)\beta(g, s) = \beta(ht, g)\beta(htg, s) = \beta(ht, g)$. Further, $\beta(g, g^{-1}) = \beta(t, gg^{-1})\beta(g, g^{-1}) = \beta(t, g)\beta(tg, g^{-1})$, so $\beta(t, g) = 1$. Hence it follows that $\beta(ht, g) = \beta(ht, g)\beta(h, t) = \beta(h, tg)\beta(t, g) = \beta(h, g)$.

Now we just need to interpret these results. We know by definition that $A[gs]$ will have in row ht an entry in place $htgs = hgs$, and that entry will be $\beta(ht, gs)$. Since this is independent of t, and lies again in the column-block for s, the blocks will indeed repeat in the style described in the theorem. Since further $\beta(ht, gs)$ is independent of s, the individual M_i are all monomial circulants coming from the twisted group algebra induced on the subgroup G of the original semigroup. □

In view of this result, we shall confine our attention from now on to algebras of monomial circulants. We can now extend Theorem 2 to monomial equivalence.

THEOREM 5. *Two algebras of monomial circulants derived from index groups G, G' and cocycles β, β' are equivalent iff there is an isomorphism $G \to G'$ that sends β to a cocycle cohomologous to β'.*

Proof. If we have an equivalence, we can decompose the monomial matrix effecting it into a product of a diagonal matrix and a permutation matrix. We have already seen that diagonal conjugation changes the cocycle to a cohomologous one for the same group. It remains only to test when a permutation matrix can conjugate one of our families to another. Conjugating the basis matrix $A[g]$ in one family by a permutation π, we get a matrix which has $A[g]_{\pi i, \pi j}$ in entry (i, j); in particular, in row 1 it has only one nonzero entry, $\beta(\pi 1, g)$ in column $\pi^{-1}(\pi(1)g)$. This matrix is in the family coming from G', and it must be $\beta(\pi 1, g)$ times $A'[\pi^{-1}(\pi(1)g)]$. Comparing the unique nonzero entries of the two in row i, we find first $\pi^{-1}(\pi(i)g) = i * \pi^{-1}(\pi(1)g)$, where $*$ is the multiplication of elements in G', and then

$$\beta(\pi i, g) = \beta(\pi 1, g)\beta'(i, \pi^{-1}(\pi(1)g)).$$

Set $\phi(g) = \pi(g)\pi(1)^{-1}$. If we let $h = \pi^{-1}(\pi(1)g)$, we have $\pi(i * h) = \pi(i)\pi(1)^{-1}\pi(h)$, or $\phi(i * h) = \phi(i)\phi(h)$; thus $\phi : G' \to G$ is an isomorphism. Let us now rewrite the β-identity by setting $t = \pi(1)$, substituting $s = \pi^{-1}(tg)$ and using ϕ in place of π; we get

$$\beta(\phi(i)t, t^{-1}\phi(s)t) = \beta(t, t^{-1}\phi(s)t)\beta'(i, s).$$

Using the cocycle identity twice, we find that $\beta(\phi(i)t, t^{-1}\phi(s)t)$ equals

$$\beta(\phi 1, t)^{-1}\beta(\phi s, t)^{-1}\beta(\phi i, \phi s)\beta(\phi(i)\phi(s), t)\beta(t, t^{-1}\phi(s)t),$$

and putting that in and cancelling we get the cohomology relation

$$\beta(\phi i, \phi s) = \gamma[i]\gamma[s]\gamma[is]^{-1}\beta'(i, s)$$

with $\gamma[g] = \beta(\phi g, t)$. □

In the twisted group algebra, change to a cohomologous cocycle replaces $A[g]$ by a scalar multiple of itself, and so we are considering two such algebras as equivalent when there is an algebra isomorphism of one to the other that respects the family of one-dimensional subspaces spanned by the group elements. This is the usual equivalence for such algebras.

3. Circulant theorems deduced from the structure of $k[G, \beta]$. We can now read off quite a good deal of information about algebras of monomial circulants from known results on twisted group algebras; the following list is only a sample. The necessary facts on cohomology and twisted group algebras are conveniently assembled in Chapters 2 and 3 of Karpilovsky's book [9], but the reader could also consult [5] or [4].

(1) A direct computation of the discriminant using the basis $A[g]$ shows that $k[G, \beta]$ is a semisimple algebra when the characteristic of k does not divide $n = |G|$. In

particular, if k is algebraically closed and char (k) does not divide n, then for our regular representation there is an invertible matrix P (not necessarily monomial) such that the matrices PAP^{-1} for A in the family have a block-diagonal structure

$$\operatorname{diag}(X_1, \cdots, X_1, X_2, \cdots, X_2, \cdots, X_m, \cdots, X_m).$$

Here X_i is an arbitrary n_i by n_i matrix repeated n_i times, and the X_i are independent of each other. In particular, $\Sigma(n_i)^2 = n$. Just as for group matrices, this tells us how the determinant for one of our algebras factors as a polynomial in the entries of the first row: there are independent linear functions $x_{pq}^{(i)}$ of the entries, falling into m families of $(n_i)^2$ each, such that the factorization of the determinant into irreducible factors is $\Pi_i(\det(x_{pq}^{(i)})^{n_i})$.

(2) If any one of the blocks X_i in (1) is 1 by 1, then the cohomology class of β is trivial, and there is a diagonal matrix which conjugates our circulants to the family of group matrices for G. In fact this is true if we simply assume (with no hypothesis on k) that there exists a common eigenvector v for the matrices in our family. Indeed, we then have $vA[g] = x(g)v$ for some nonzero scalar $x(g)$, and the multiplication rule (**) gives $x[g]x[p] = \beta(g, p)x[gp]$. But this says that β is cohomologous to the trivial cocycle.

(3) It is clear from (**) that the matrices in the algebra all commute with each other iff (a) the group G is abelian, and (b) the cocycle β is symmetric (i.e., $\beta(p, q) = \beta(q, p)$ for all p, q in G). If in addition the field k is algebraically closed of characteristic not dividing n, then the algebra will necessarily be conjugate to the algebra of all diagonal n by n matrices. The monomial equivalence families of this type correspond precisely to the different abelian groups of order n. This follows from (2) above, which implies that here all cohomology classes are trivial.

(4) Suppose that k is algebraically closed, or more precisely just that it is closed under taking nth roots. If μ_n denotes the nth roots of 1 in k, we have then the exact sequence of multiplicative groups

$$1 \to \mu_n \to k^* \xrightarrow{n} k^* \to 1,$$

and in turn by [9, p. 42] this gives us a cohomology exact sequence

$$1 \to H^1(G, k^*) \to H^2(G, \mu_n) \to H^2(G, k^*) \to 1.$$

In particular, all classes in $H^2(G, k^*)$ are represented by cocycles with values in μ_n. In terms of the earlier λ-group matrices, this says that (under our hypothesis on k) any family of them can be conjugated by diagonal matrices to one where the kernel H is cyclic of order n. As $H^1(G, k^*) = \operatorname{Hom}(G, k^*) = \operatorname{Hom}(G, \mu_n)$ is computable, the sequence also allows us to compute the size of $H^2(G, k^*)$ in finitely many steps.

(5) In particular, if k satisfies the hypothesis in (4), then there are only finitely many different equivalence classes of algebras of n by n monomial circulants. For there are only finitely many ways to make the index set into a group; and for each group G there are only finitely many cocycles with values in μ_n, and hence each $H^2(G, k^*)$ is finite. Theorem 4 then shows more generally that there are only finitely many equivalence classes closed under multiplication.

A similar result holds when k is the field $\mathbb{R}$ for real numbers. For here we can write $\mathbb{R}^*$ as a direct product $\mu_2 \times \mathbb{R}_{>0}$, and the multiplicative group of positive reals is uniquely divisible by n. Hence the cohomology with values in it is trivial, and we have $H^2(G, \mathbb{R}^*) \cong H^2(G, \mu_2)$ for any G. Thus we get the following result, which is not at all obvious from the original definitions:

THEOREM 6. *Over either the real numbers or the complex numbers, there are only finitely many different classes of n by n monomial circulants closed under multiplication. For any given n, representatives for all the classes can be computed.* □

(6) On the other hand, for fields like $k = \mathbb{Q}$ there will be infinitely many different types. Indeed, suppose that we take G cyclic of order n (where the trivial cocycle yields ordinary circulants). Then $H^2(G, k^*)$ is isomorphic to $k^*/(k^*)^n$, and for $k = \mathbb{Q}$ this quotient is infinite for every $n \geqq 2$. If g is a generator of G, we get a cocycle corresponding to a given element r in k^* by letting

$$\begin{aligned} \beta(g^i, g^j) &= 1 \quad \text{if } i+j<n, \\ &= r \quad \text{if } i+j \geqq n. \end{aligned}$$

Here $A[g]^n = rI$, and the algebra is isomorphic to $k[X]/(X^n - r)$. One familiar example of this is the family of "skew circulants," where $r = -1$.

(7) Generalizing (3) above, one can directly compute which matrices $\Sigma c[g]A[g]$ in an algebra of monomial circulants commute with the whole algebra. The result is that each $c[g]$ determines the value of all $c[hgh^{-1}]$, and also $c[g]$ must actually be zero unless the cocycle β satisfies

$$\beta(g,p) = \beta(p,g) \quad \text{for all } p \text{ commuting with } g.$$

Thus the dimension of the center is equal to the number of conjugacy classes in G for which this condition on the cocycle is satisfied, and we can explicitly determine a basis for the center. In the situation of (1), the dimension of the center tells us the number of blocks X_i. In particular, the number of blocks is never greater than the number of conjugacy classes in G. We can find the sizes of the blocks by diagonalizing the basis elements of the center.

4. Examples for order 4. Over any field there are at least two equivalence classes of 4 by 4 algebras of monomial circulants, one for each group of order 4; the corresponding matrices are

$$\begin{bmatrix} a & b & c & d \\ d & a & b & c \\ c & d & a & b \\ b & c & d & a \end{bmatrix},$$

the ordinary circulants corresponding to the cyclic group of order 4, and

$$\begin{bmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{bmatrix}$$

corresponding to the Klein 4-group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Over a field like $\mathbb{Q}$, there are infinitely many others.

When $k = \mathbb{C}$, we can compute all equivalence classes as in (5) in § 3. There are no additional types arising from the cyclic group, because $H^2(\mathbb{Z}/4\mathbb{Z}, \mathbb{C}^*) = \mathbb{C}^*/(\mathbb{C}^*)^4 = 1$. When $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, multiplication by 2 is a homomorphism that annihilates G and hence annihilates all cohomology of G, and thus in the reasoning of (4) in § 3 we can replace μ_4 by $\mu_2 = \{\pm 1\}$. Mechanical computation shows that there are exactly 16 (normalized) cocycles β on G with values in μ_2, and there is just one nontrivial coboundary to change them, so there are eight elements in $H^2(G, \mu_2)$. But there are four elements in

$H^1(G, \mathbb{C}^*)$, and hence there are two elements in $H^2(G, \mathbb{C}^*)$. The nontrivial one (which is of course supplied to us by the computation) can be represented by matrices of the following form:

$$A=\begin{bmatrix} a & b & c & d \\ -b & a & -d & c \\ c & -d & a & -b \\ d & c & b & a \end{bmatrix}.$$

The center contains only multiples of the identity, and hence our algebra must be abstractly isomorphic to 2 by 2 matrices. Explicitly, set

$$Q=\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix}.$$

Then it is easy to compute that the general matrix A above satisfies

$$QAQ^{-1}=\begin{bmatrix} a+c & b-d & 0 & 0 \\ -b-d & a-c & 0 & 0 \\ 0 & 0 & a+c & b-d \\ 0 & 0 & -b-d & a-c \end{bmatrix}.$$

Notice that (up to a scalar factor) the matrix Q that we took here is *unitary*. In the next section we shall show that we can do almost as well in all cases over $\mathbb{C}$. The only other thing we may need is a diagonal factor to adjust for the absolute values of β.

We can extend this example [9, p. 62] to compute $H^2(G, \mathbb{C}^*)$ for all abelian groups G, and a little attention to the details of the proof will produce the cocycles (and thus the circulant families) explicitly.

Now consider the possible structures over $\mathbb{R}$. As we saw in (5) above, $H^2(G, \mathbb{R}^*) \cong H^2(G, \mu_2)$. For G cyclic of order 4, we get two possibilities. One type is again the usual circulant matrices corresponding to the usual group algebra; the other is the "skew-circulant" matrices

$$\begin{bmatrix} a & b & c & d \\ -d & a & b & c \\ -c & -d & a & b \\ -b & -c & -d & a \end{bmatrix}$$

given by the formula in (6) in § 3 with $r = -1$.

For $G \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, we have (as we saw before) eight elements in $H^2(G, \mu_2)$. But this does *not* directly tell us the number of monomial equivalence classes, because in Theorem 5 it is possible to have $G = G'$ with a nontrivial map between them. That is, the automorphism group Aut (G) acts on $H^2(G, k^*)$, and two classes in the same orbit of that action will give algebras differing only by permutations of the basis elements. In our particular case, Aut (G) has six elements. We know already that four of the eight classes in $H^2(G, \mathbb{R}^*)$ will become trivial in $H^2(G, \mathbb{C}^*)$, and all these correspond to symmetric cocycles. It is easy to calculate that the three nontrivial symmetric classes form a single orbit for Aut (G) and give us just one more type of generalized circulant, which we can write as

$$\begin{bmatrix} a & b & c & d \\ b & a & d & c \\ -c & -d & a & b \\ -d & -c & b & a \end{bmatrix}.$$

By (2) in § 3 we know in advance that this is a commutative semisimple algebra of real matrices having no common real eigenvector, and hence it is isomorphic as an algebra to $\mathbb{C} \times \mathbb{C}$.

The other four cohomology classes must be in the other coset of the image of $H^1(G, \mathbb{C}^*)$; they yield the nontrivial class over $\mathbb{C}$ and are nonsymmetric. They again fall into just two orbits, one being again the type isomorphic to the 2 by 2 matrix algebra. The other can be written for instance as follows:

$$\begin{bmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{bmatrix}.$$

This is the right regular representation for the quaternions (in basis 1, i, j, k). In rough summary, we have the following theorem.

THEOREM 7. *There are precisely six equivalence classes of real* 4 *by* 4 *algebras of monomial circulants. Over the complex numbers, they collapse to three classes.* □

5. Unitary reduction over $\mathbb{C}$. In [11], [12], [13] it has been shown (by a computation using eigenvalues) that any commutative algebra of monomial circulants over $\mathbb{C}$ can be reduced to diagonal form under conjugation by the product of a unitary matrix and a diagonal matrix; this is a stronger statement than the abstract result that follows from (1) in § 3. We saw further in § 4 that in a particular (4 by 4) noncommutative case we could likewise reduce the family to its standard block form by such a transformation. In this section we establish that result in general.

THEOREM 8. *For any algebra of monomial circulants over* $\mathbb{C}$, *there exist a diagonal matrix D* (*with real positive entries*) *and a unitary matrix U such that conjugation by UD puts the algebra into the block form of* (1) *in* § 3.

Proof. We begin with an explicit version of the reduction done abstractly in (4) in § 3. The cocycle identity on β tells us that

$$\beta(c, d)\beta(cd, g) = \beta(c, dg)\beta(d, g).$$

If we define $\gamma(d)$ to be the product of all $\beta(c, d)$, then we find by taking products over c that

$$\gamma(d)\gamma(g) = \gamma(dg)\beta(d, g)^n.$$

Hence the absolute values of the entries satisfy

$$|\beta(d, g)| = |\gamma(d)|^{1/n}|\gamma(g)|^{1/n}/|\gamma(dg)|^{1/n}.$$

Since the cocycle is normalized, we have of course $\gamma(1) = 1$. If we let D be the diagonal matrix diag $(|\gamma(g)|^{-1/n})$ and replace $A[g]$ by $DA[g]D^{-1}$ as in the proof of Theorem 4, we get an algebra corresponding to a cocycle β' for which all $|\beta'(g, p)| = 1$. A straightforward computation now establishes the following lemma.

LEMMA 9. *When* $|\beta'| = 1$, *the conjugate transpose matrix* $A[g]^*$ *is equal to* $\beta'(g, g^{-1})^{-1}A[g^{-1}]$. *In particular, when* $|\beta'| = 1$, *the algebra of monomial circulants is closed under taking conjugate transposes.* □

Algebras of matrices with this selfadjointness property are known to be quite special, essentially because the orthogonal complement of any invariant subspace is again an invariant subspace. Specifically [10, pp. 5–8], under conjugation by a unitary matrix, any such algebra can be transformed to one of the following form:

$$\text{block diag}\,(X_1, \cdots, X_1, X_2, \cdots, X_2, \cdots, X_m, \cdots, X_m)$$

as in (1) of § 3, except that in general here the number of repetitions of each independent block X_i can depend on the particular algebra involved. But we know already that we are dealing with a regular representation, and hence an n_i by n_i block must occur n_i times whenever we get the matrices into block form. Thus Theorem 8 is proved. □

COROLLARY 10. *Let $x_{pq}^{(i)}$ be a list (in some order) of the entries in the block matrix above. Then there is a unitary map that sends the n-tuple $(|\gamma(g)|^{1/n}c(g))$ to the corresponding n-tuple $((n_i/n)^{1/2}x_{pq}^{(i)})$.*

Proof. The first n-tuple is the first row of $\Sigma c(g)DA[g]D^{-1}$ above, so we may make that change first and assume that our cocycle has absolute value 1. Now on n by n matrices we have an inner product $(1/n)$ Tr (AB^*), for which the matrices with single nonzero entries are orthogonal. By Lemma 9, the $A[g]$ are orthonormal (and the map $A \mapsto e_1A$ is an isometry from our algebra to $\mathbb{C}^n$). The map $B \mapsto UBU^{-1}$ for unitary U is an isometry on the whole matrix algebra, and thus the $UA[g]U^{-1}$ are an orthonormal basis for the block-form image. But the matrices $E_{pq}^{(i)}$ with one block entry equal to 1 are also an orthogonal basis. As the block X_i is repeated n_i times, the norm of $E_{pq}^{(i)}$ in our inner product is $(n_i/n)^{1/2}$. Thus $(n/n_i)^{1/2}E_{pq}^{(i)}$ is another orthonormal basis for our block matrix algebra, and the transformation between coordinates in the two bases is unitary. □

6. Monomial circulants preserved by the algebras. It is known [14] that the g-circulants are characterized among all permutation circulants by the fact that they are preserved under multiplication by ordinary circulants. To find their analogues, then, we must determine the families of monomial circulants preserved under multiplication by some algebras of monomial circulants. We shall first give the result of a direct analysis and then show how the families are related to endomorphisms of twisted group algebras.

We start with a fixed algebra of monomial circulants coming from a group G and a cocycle β. As before, we denote the basis matrices by $A[g]$. Consider now some other monomial circulants, with permutation $p \mapsto i*p$ in row i and nonzero scalars $\zeta(i, p)$; recall that by definition $1*p = p$ and $\zeta(1, p) = 1$. A basis of the family is then given by the matrices $M[r]$ with nonzero entries $\zeta(i, r)$ in row i and column $i*r$. A computation very much like those that have gone before establishes the following criterion.

THEOREM 11. *The monomial circulants spanned by the $M[r]$ are preserved under multiplication by the $A[g]$ (on both sides) iff*

(1) *$i*r = \phi(i)r$ for some homomorphism $\phi : G \to G$, and*

(2) *the $\zeta(i, j)$ satisfy the identities*

$$\zeta(i, g) = \zeta(i, 1)\beta(\phi(i), g)$$

and

$$\zeta(ig, 1)\beta(1, g) = \zeta(g, 1)\zeta(i, 1)\beta(\phi(i), \phi(g)). \qquad \square$$

Re-interpreting condition (2), we see that monomial circulants of the type in Theorem 11 exist for given G and β and a given homomorphism $\phi : G \to G$ iff the cocycle $\beta\circ(\phi, \phi)$ is cohomologous to β, or, in other words, iff the cohomology class of β is preserved by ϕ.

COROLLARY 12. *There are only finitely many families of monomial circulants preserved under (two-sided) multiplication by the circulant algebra corresponding to a fixed G and β.*

Proof. There are only finitely many possible $\phi : G \to G$. For those that satisfy the necessary condition, the choice of the values $\zeta(g, 1)$ (which determine all others) is unique up to multiplication by values $\gamma(g)$ satisfying $\gamma(ig) = \gamma(i)\gamma(g)$. Such γ form a homomorphism from G to k^*, and there are only finitely many such homomorphisms. □

If we change the algebra of monomial circulants by a (monomial) equivalence, then we change the circulant families preserved by it in the same way. In view of Theorem 6, then, we have the following result:

THEOREM 13. *Over the real or complex numbers, there are only finitely many different equivalence classes of monomial circulants of a given size that are preserved under multiplication by some algebra of monomial circulants. In principle, we can compute them all.* □

Now we can connect these families with our twisted group algebras. Just as we interpreted the $A[g]$ as right multiplication maps on $k[G, \beta]$, so now we shall interpret the new families as linear transformations on that algebra, using as before $A[g]$ as the element corresponding to the basis vector e_g of k^n.

THEOREM 14. *Each family of monomial circulants preserved by a given algebra of monomial circulants is given (as linear transformations on $k[G, \beta]$) by maps of the form $x \mapsto \Phi(x)v$ for varying v in $k[G, \beta]$. Here Φ is an algebra endomorphism of $k[G, \beta]$ that sends each $A[g]$ to a multiple of some $A[h]$. Conversely, every such Φ gives such a family of circulants.*

Proof. It is trivial to compute that a monomial mapping Φ sending $\Sigma c(g)A[g]$ to $\Sigma c(g)A[\phi(g)]y(g)$ will preserve multiplication iff ϕ is a homomorphism and the $y(g)$ satisfy $y(g)y(h)\beta(\phi g, \phi h) = y(gh)\beta(g,h)$, precisely the conditions required for $\zeta(g, 1)$ in Theorem 11. The condition $y(1) = 1$ is implied because we must preserve the unit element, and then we get $y(g) \neq 0$ by taking $h = g^{-1}$ in the equations. We then just compute that the mapping sending $A[i]$ to $\Phi(A[i])A[r]$ for some fixed $A[r]$ has image $y(i)\beta(\phi(i), r)A[\phi(i)r]$, which agrees exactly with the formula for $\zeta(i, r)$ in Theorem 11. □

This result says that the circulant matrices in question are very close to the regular representation, differing from it only in that some fixed algebra endomorphism is applied before the right multiplication. In particular, we have the following fact.

COROLLARY 15. *The family of matrices in Theorem* 11 *contains an invertible matrix iff the endomorphism ϕ is an automorphism.*

Proof. Clearly Φ is bijective iff ϕ is. When this is so, then actually $M[1]$ (corresponding to $v = 1$ in the above formula) is an invertible mapping. When Φ is not bijective, every element in its kernel is in the nullspace of all matrices in the family. □

In view of Theorem 14, we shall call the monomial circulant families occurring in Theorem 11 **monomial endomorphism circulants;** those containing an invertible matrix (which, as with the algebras, are likely to be the most important) will be **monomial automorphism circulants.** Since $\det(x \mapsto \Phi(x)v) = \det(\Phi) \det(x \mapsto xv)$, we have the following factorization.

COROLLARY 16. *The determinant on matrices in a family of monomial automorphism circulants is a polynomial in the first-row entries with a factorization of the same kind as that for the algebra.* □

We should also note how matrix multiplication of such families corresponds to composition.

PROPOSITION 17. *Let $\mathcal{M}$ and $\mathcal{N}$ be two families of monomial endomorphism circulants for the same algebra. Let Φ, Ψ be the corresponding endomorphisms. Then the matrix products in $\mathcal{M}\mathcal{N}$ are the family of monomial endomorphism circulants for that same algebra and the endomorphism $\Psi \circ \Phi$.*

Proof. Each matrix product sends a vector x to an image of the form $\Psi[\Phi(x)v]w$, which equals $[\Psi \circ \Phi(x)][\Psi(v)w]$. □

This is a wide generalization of the familiar fact [1] that the product of a g-circulant and an h-circulant is a gh-circulant.

7. Equivalence classes of monomial automorphism circulants. As with algebras of circulants, we can find an algebraic classification of the families we have just studied. Again we focus only on those containing invertible matrices.

LEMMA 18. *A family of monomial automorphism circulants uniquely determines the algebra of monomial circulants from which it came.*

Proof. The monomial automorphism family has dimension n and contains some invertible matrix M. The map $X \mapsto MX$ on matrices is bijective, and hence at most an n-dimensional family of matrices X can send this particular M into the family under right multiplication. Since by definition the circulant algebra does this, it coincides with the family of all such matrices and is thereby determined. □

If two families of monomial automorphism circulants are equivalent, then the same monomial conjugation will give an equivalence of the associated algebras of monomial circulants. Thus to determine the equivalence classes it will be enough to determine those arising for a fixed algebra. To express the classification, let M Aut $(k[G, \beta])$ be the group of all monomial automorphisms of $k[G, \beta]$. Straightforward computation shows that there is an exact sequence of groups as follows:

$$1 \to \mathrm{Hom}\,(G, k^*) \to M\,\mathrm{Aut}\,(k[G, \beta]) \to \mathrm{Aut}_\beta\,(G) \to 1,$$

where $\mathrm{Aut}_\beta\,(G)$ is the subgroup of Aut (G) preserving the class of β in $H^2(G, k^*)$ and the induced action of this group on Hom (G, k^*) is the natural one.

THEOREM 19. *Consider two families of monomial automorphism circulants corresponding to elements Φ, Ψ of M* Aut $(k[G, \beta])$ *for the same algebra of monomial circulants. Then they are equivalent iff Φ and Ψ are conjugate in M* Aut $(k[G, \beta])$.

Proof. Let us suppose first that the two families of matrices are conjugate by a monomial transformation. We view both families as acting on $k[G, \beta]$, and we let $F: k[G, \beta] \to k[G, \beta]$ be the monomial map (written on the left) that conjugates one to the other. One of the families comprises the maps sending the element x to $\Phi(x)u$ for various u; the other comprises the maps sending x to various $\Psi(x)v$. The assumption thus is that for each u in the algebra there is some other v such that $\Phi(Fx)u = F(\Psi(x)v)$ for all x. Setting $x = 1$ $[=e_1]$, we have $\Phi(F1)u = F(v)$, and this equation determines u from v. Putting in that value, we get the equation

$$\Phi(Fx)\Phi(F1)^{-1}F(v) = F(\Psi(x)v),$$

for all x and v. Setting $v = 1$, we can solve to find $\Phi(Fx)\Phi(F1)^{-1}F(1) = F(\Psi x)$, and then we can put in the resulting value for $\Phi(Fx)$ to get

$$F(\Psi x)F(1)^{-1}\Phi(F1)\Phi(F1)^{-1}F(v) = F((\Psi x)v).$$

Now let $H(x) = F(1)^{-1}F(x)$. Then the condition says that $H(\Psi x)H(v) = H((\Psi x)v)$; as Ψ is bijective, H is an automorphism of the algebra. Since F is monomial, $F(1)$ is a scalar times a basis element of $k[G, \beta]$; and since the multiplication by these basis elements is monomial, the automorphism H is in M Aut $(k[G, \beta])$.

Now let $r = F(1)$. If we rewrite our basic identity in terms of H and r, with $F(x) = rH(x)$, we get

$$\Phi(rH(X))\Phi(r)^{-1}rH(v) = rH(\Psi(x)v) = rH(\Psi x)H(v).$$

If we here put $v = 1$, we see that this is equivalent to

$$\Phi(r)\Phi(H(x))\Phi(r)^{-1} = rH(\Psi x)r^{-1}.$$

Let $\Lambda(x) = rH(x)r^{-1}$. Since $r = F1$ is a scalar times a basis element, this operation Λ is also in M Aut $(k[G, \beta])$. Since F is an automorphism, our equation then tells us that

$\Phi\Lambda = \Lambda\Psi$, or in other terms $\Lambda^{-1}\Phi\Lambda = \Psi$, as claimed. Conversely, we can read the whole computation backwards (taking $r = 1$ and $F = \Lambda$) to show that this condition is sufficient. □

For example, let us return to the 4 by 4 case over $\mathbb{C}$. We found the three inequivalent monomial circulant algebras in § 4. For a group algebra (trivial cohomology class), all elements of Aut (G) preserve the class. Furthermore, a group automorphism cannot take a nontrivial cohomology class to a trivial one, so the unique algebra class for $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ with nontrivial cocycle will also have its class preserved by all automorphisms. For group algebras, there is always a subgroup of M Aut $(k[G])$ naturally isomorphic to Aut (G); thus the exact sequence mentioned at the start of the section splits, and we have a semidirect product (and the monomial automorphism circulants coming from the elements in Aut (G) are actually permutation circulants). For $G = \mathbb{Z}/4\mathbb{Z}$, it is easy to compute that M Aut $(k[G])$ is the dihedral group of order 8, and hence there are five conjugacy classes in it, giving rise to five equivalence classes of monomial automorphism circulants. The conjugacy class consisting of the identity gives the algebra itself. A typical example of the other classes is given by the family of matrices

$$\begin{bmatrix} a & b & c & d \\ -b & -c & -d & -a \\ c & d & a & b \\ -d & -a & -b & -c \end{bmatrix}.$$

Like g-circulants, this family will be closed under multiplication on both sides by ordinary 4 by 4 circulants. As was predicted by Corollary 16, the determinant here factors in the same way as for ordinary circulants.

For the trivial cocycle on $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, we have M Aut isomorphic to the alternating group A_4. Because $|G|$ is relatively prime to $|\text{Aut}\,(G)|$, the group M Aut for the nontrivial class also splits to give the semidirect product A_4. There are 4 conjugacy classes in A_4, and thus this G yields eight equivalence classes of monomial automorphism circulants (two of them algebras). Thus

COROLLARY 20. *There are precisely* 13 *equivalence classes of complex* 4 *by* 4 *monomial automorphism circulants.* □

Those who need to compute more general examples may want to recall one standard theorem that did not arise in these abelian examples: the inner automorphisms of G are always contained in $\text{Aut}_\beta\,(G)$. This is actually obvious with our algebras, since the inner automorphism $x \mapsto A[g]^{-1}xA[g]$ on the twisted group algebra induces the inner automorphism $h \mapsto g^{-1}hg$ in Aut (G).

Appendix. Linear circulants. It is clearly possible to replace monomial transformations by various other linear groups of transformations. In this Appendix we look very quickly at the most general possibility, where the entries in subsequent rows are obtained as (invertible) linear transformations of the first row. More precisely, we begin with a fixed sequence of invertible matrices $C = \{C[1], \cdots, C[n]\}$ with $C[1] = I$; this will define our style of circulants. Starting then with any row u in k^n (basis vectors e_i), we get a matrix $M(u)$ $\{$or $M_C(u)$, if we must be precise$\}$ with row i given by $uC[i]$. Clearly the family $\mathcal{M}$ of all such matrices is n-dimensional, and we call it the family of **linear circulants** defined by C. It has the additional property that its individual rows are arbitrary, i.e., $e_i\mathcal{M} = k^n$ for each i; conversely, a simple dimension count shows that an n-dimensional space of n by n matrices with this property is indeed a family of linear circulants. Thus we are dealing with fairly general families. Again we focus on those that are closed under multiplication.

THEOREM A1. (a) *The linear circulant family $\mathcal{M}$ is closed under multiplication iff there is an associative bilinear multiplication $*$ on k^n for which e_1 is a left unit and $uC[i] = e_i * u$ for all i. In that case, we have in general $wM(u) = w*u$.*

(b) *In the situation of* (a), *the following are equivalent*:

(i) *One of the $M(u)$ is invertible.*

(ii) *The identity matrix is in $\mathcal{M}$.*

(iii) *e_1 is a two-sided unit in the algebra.*

(iv) *For every i, the first row of $C[i]$ is e_1.*

When these conditions hold, then the inverse of every invertible matrix in $\mathcal{M}$ is again in $\mathcal{M}$.

Proof. (a) If we have the multiplication on k^n, we define $C[i]$ by the condition that row j in it be $e_i * e_j$. Bilinearity then shows that $uC[i] = e_i * u$, and so more generally $wM(u) = w*u$. By associativity then $wM(u)M(v) = (w*u)*v = w*(u*v) = wM(u*v)$, and the family of matrices is closed under multiplication. The condition on e_1 tells us that $C[1]$ is the identity.

Conversely, if we have a family closed under multiplication, the condition on $C[1]$ tells us that $e_1 M(v) = v$, so that $v \mapsto M(v)$ is a linear bijection from k^n to $\mathcal{M}$. As $\mathcal{M}$ has an associative bilinear multiplication, we can pull it back to get such a multiplication on k^n; specifically, if $M(u)M(v) = M(z)$, then $z = e_1 M(z) = e_1 M(u)M(v) = uM(v)$, so we see that the multiplication $*$ is indeed given by the formula in the theorem. By construction also e_1 is a left identity.

Before continuing with part (b), we should insert a lemma analyzing the structure possible at this stage. The older papers on this topic [7], [2] do not seem to contain precisely the result we need, but it is really just the Peirce decomposition for the idempotent e_1 and thus little more than an exercise.

LEMMA A2. *Let R be a ring with a left identity, e. Then R has the form $R_0 \times M$, where R_0 is a subring having e as a two-sided identity and M is a unital left R_0-module; the multiplication has the form*

$$(r,m)*(r',m') = (rr', rm').$$

Conversely, any such construction yields a ring with a left identity. If M is nonzero (*e not a two-sided identity*), *then there are other left identities; they all yield the same M, while R_0 is replaced by $\{(r, rm) | r \in R_0\}$ for an m in M that is* (*fixed but*) *arbitrary. In particular, R_0 and M are determined up to isomorphism by the original R.* □

Now we prove part (b) of Theorem A1. Let $M(u)$ be an invertible element in $\mathcal{M}$. Then all sums of powers of $M(u)$ are in $\mathcal{M}$. The characteristic equation has nonzero constant term, and so we can express I in terms of these powers; hence I is in $\mathcal{M}$. (The converse implication is evident.) As there is thus a two-sided identity in $\mathcal{M}$, it must coincide with the left identity given in the algebra by e_1. Conversely, if e_1 is a two-sided identity, then $u = u*e_1 = uM(e_1)$, so $M(e_1) = I$. We have $e_i = e_i * e_1 = e_1 C[i]$, so we find that condition (iv) is also equivalent. Finally, if $M(v)$ is invertible, then its inverse is a linear combination of I and powers of $M(v)$, and hence it is again in $\mathcal{M}$. □

We can of course write out conditions on the $C[i]$ that yield various properties of the algebra. Here for instance is the basic result on diagonalization.

PROPOSITION A3. *Let $\mathcal{M}$ be a family of linear circulants over a field k. The following are equivalent*:

(a) *$\mathcal{M}$ is an algebra, and each matrix in it is diagonalizable over k;*

(b) *$P\mathcal{M}P^{-1}$ consists of diagonal matrices for some matrix P.*

If the field k is algebraically closed, these are equivalent to the following identities:

(c1) $\Sigma_r C[i]_{jr} C[r]_{pt} = \Sigma_r C[j]_{pr} C[i]_{rt}$;

(c2) $e_1 C[i] = e_i$;

(c3) $e_i C[j] = e_j C[i]$;

(c4) $0 \neq \det_{ij} (\Sigma_{k,r} C[k]_{ir} C[r]_{jk})$.

Proof. Clearly (b) implies (a). Conversely, if we have (a), we know the matrices are the right regular representation of some algebra. By assumption, none of the matrices can be nilpotent. Hence the algebra is semisimple. Thus it is a product of matrix algebras over division rings. These matrices must be 1 by 1, since otherwise we would again have nilpotent elements. Thus we have a product of division rings. If any one of them is not $= k$, then the regular representation of something in it outside k is not diagonalizable over k. Thus we have the regular representation of k^n, and in an appropriate basis that will coincide with the family of diagonal matrices. In (c), condition (1) makes the family closed under multiplication, condition (2) gives it an identity, condition (3) then makes it commutative, and condition (4) then makes it separable [3, p. 45] by making its discriminant nonzero, $0 \neq \det(\mathrm{Tr}\,(e_i * e_j))$. □

To conclude, we briefly discuss the families preserved by such algebras. The proof of Theorem 14 shows in this context that if $\mathscr{A}$ is a linear circulant algebra and $\mathscr{M}$ is a linear circulant family with $\mathscr{M}\mathscr{A} = \mathscr{M} = \mathscr{A}\mathscr{M}$, then the matrices in $\mathscr{M}$ are those corresponding to maps $u \mapsto \Phi(u) * v$ for some endomorphism Φ of the algebra $\mathscr{A}$. The family includes invertible matrices iff Φ is an automorphism. If $\mathscr{N}$ is another family corresponding to $\mathscr{A}$ and the endomorphism Ψ, then as in Proposition 17 we find that the matrix products in $\mathscr{M}\mathscr{N}$ are a family corresponding to $\mathscr{A}$ and the endomorphism $\Psi\Phi$.

The intrinsic definition of the algebra from the matrices in Theorem A1 shows at once that two linear circulant families are conjugate by an invertible matrix iff the algebras involved are conjugate. As in Lemma 18, a family of linear automorphism circulants determines the associated algebra; and as in Theorem 19, we find that two of them for the same algebra are conjugate iff the automorphisms are conjugate in Aut $(\mathscr{A})$. Finally we have the following result, one that shows how our generalized circulant algebras form a reasonably self-contained topic.

THEOREM A4. *Let $\mathscr{M}$ be a family of linear circulants containing an invertible matrix. Suppose there is some n-dimensional subalgebra $\mathscr{A}$ of matrices for which $\mathscr{A}\mathscr{M} = \mathscr{M}$. Then in fact $\mathscr{A}$ is a linear circulant algebra.*

Proof. Let $C[i]$ be the matrices defining $\mathscr{M}$. By assumption, there is some vector u with the $uC[i]$ all independent. For each r, then, we can write $uC[i]C[r] = \Sigma_s D[r]_{is} uC[s]$; that is, $D[r]$ is the matrix expressing multiplication by $C[r]$ in this new basis. In particular, $D[r]$ is invertible and $D[1]$ is the identity. Now for each A in $\mathscr{A}$, $AM(u)$ is contained in $\mathscr{M}$, say $= M(v)$. The first row of $M(v)$ is v, and so we have $v_r = \Sigma_k A_{1k}(uC[k])_r$. Then row i of $M(v)$ will be $vC[i]$, and this must agree with row i of $AM(u)$, with entry in column s given by $\Sigma_t A_{it}(uC[t])_s$. Hence this latter expression is equal to $\Sigma_{r,k} A_{1k}(uC[k]_r) C[i]_{rs}$. Now this can be rewritten as

$$\Sigma_k A_{1k}(uC[k]C[i])_s = \Sigma_k A_{1k} \Sigma_t D[i]_{kt}(uC[t])_s.$$

Thus we have the vector equality $\Sigma_t A_{it}(uC[t]) = \Sigma_t(\Sigma_k A_{1k} D[i]_{kt})(uC[t])$. Since the $uC[t]$ are independent, we conclude that

$$A_{it} = \Sigma_k A_{1k} D[i]_{kt}.$$

Thus indeed the matrices A are in a fixed linear circulant family. □

REFERENCES

[1] C. M. Ablow and J. C. Brenner, *Roots and canonical forms for circulant matrices*, Trans. Amer. Math. Soc., 107 (1963), pp. 360–376.

[2] R. Baer, *Inverses and zero-divisors*, Bull. Amer. Math. Soc., 48 (1942), pp. 630–638.

[3] N. Bourbaki, *Algèbre* 9: *Formes sesquilinéaires et formes quadratiques*, Hermann, Paris, 1959.

[4] S. B. Conlon, *Twisted group algebras and their representations*, J. Austral. Math. Soc., 4 (1964), pp. 152–173.

[5] C. W. Curtis and I. Reiner, *Methods of Representation Theory*, Vol. 1, John Wiley, New York, 1981.

[6] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1977.

[7] C. Hopkins, *Rings with minimal condition for left ideals*, Ann. of Math., 40 (1939), pp. 712–730.

[8] J. M. Howie, *An Introduction to Semigroup Theory*, Academic Press, London, 1976.

[9] G. Karpilovsky, *Projective Representations of Finite Groups*, Marcel Dekker, New York, 1985.

[10] J. T. Schwartz, *W^*-Algebras*, Gordon and Breach, New York, 1967.

[11] K. Wang, *Theory of generalized group matrices*, J. Algebra, 76 (1982), pp. 153–170.

[12] K. Wang and P. J. Davis, *Permutations and group matrices*, preprint.

[13] ———, *Theory of λ-group matrices*, preprint.

[14] W. C. Waterhouse, *Circulant-style matrices closed under multiplication*, Linear and Multilinear Algebra, 18 (1985), pp. 197–206.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

015

INVERTIBLE SELFADJOINT EXTENSIONS OF BAND MATRICES AND THEIR ENTROPY*

ROBERT L. ELLIS†, ISRAEL GOHBERG‡ AND DAVID C. LAY†

Abstract. A maximum entropy principle for positive definite extensions of band matrices is generalized here to a large class of indefinite selfadjoint matrices. It is known that a selfadjoint band matrix R with certain nonvanishing minor determinants has a unique extension to an invertible selfadjoint matrix F such that F^{-1} is a band matrix. Sufficient conditions are described here such that $|\det F| > |\det G|$ when G is any other invertible selfadjoint extension of F.

Key words. selfadjoint, hermitian, Toeplitz matrix, entropy, extension

AMS(MOS) subject classifications. 47A20, 47A68

Introduction. The following problem was considered in [5]. Let m and n be integers with $0 \leqq m \leqq n-2$. Suppose the entries of an $n \times n$ selfadjoint matrix F are specified only for $|k-j| \leqq m$ (that is, in a "band" with "bandwidth" m). Under what conditions is it possible for F to be positive definite with all entries of F^{-1} equal to 0 outside the band of F^{-1} with bandwidth m? This problem has important connections with signal processing, system theory and other areas. It can be reformulated as follows. We say that an $n \times n$ matrix $R = (R_{jk})$ is an m-band matrix if $R_{jk} = 0$ for $|k-j| > m$, and that an $n \times n$ matrix F extends such an m-band matrix R if $F_{jk} = R_{jk}$ for $|k-j| \leqq m$. Then the problem is to determine conditions on an m-band matrix R under which R has a positive definite extension whose inverse is an m-band matrix. Throughout the paper, if M is a matrix, then $M(j, \cdots, k)$ denotes the principal submatrix $(m_{pq})_{j \leqq p,q \leqq k}$. In [5, Thm. 6.1] it was proved that if $R(j, \cdots, j+m)$ is positive definite for $1 \leqq j \leqq n-m$, then there is a unique positive definite extension F of R whose inverse is an m-band matrix. (See also [1].) This so-called band extension F can also be characterized as the unique positive definite matrix whose determinant is as large as possible. It was also proved that F is Toeplitz if R is.

In [6], a "permanence principle" was found which states that if F is the band extension of an m-band matrix R satisfying the conditions above, then any principal submatrix $F(j, \cdots, k)$ of F is the band extension of the corresponding principal submatrix $R(j, \cdots, k)$ of R. This principle implies that the band extension F can be obtained by a series of "one-step" extensions of band matrices each having bandwidth two less than its size. A selfadjoint one-step extension F of an $n \times n$ band matrix R is determined by its $(n, 1)$-entry w. In [6] it was shown that the set of all w for which F is positive definite is the interior of a disk whose center is the unique value of w for which F is the band extension of R.

The main purpose of this paper is to generalize the preceding results to indefinite selfadjoint matrices. To simplify the exposition we assume the bandwidth m is positive, but our results extend easily to the case $m = 0$. The positivity assumptions on submatrices of R will be replaced by the requirement that the matrices $R(j, \cdots, j+m)$ be invertible for $1 \leqq j \leqq n-m$ and the matrices $R(j+1, \cdots, j+m)$ be invertible for $1 \leqq j \leqq n-m-1$. The results depend on the signs of the following determinants:

$$D_k = \det R(k, \cdots, k+m) \qquad (1 \leqq k \leqq n-m)$$

* Received by the editors December 29, 1985; accepted for publication December 24, 1986.

† Department of Mathematics, University of Maryland, College Park, Maryland 20742.

‡ School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel.

and

$$d_k = \det R(k+1, \cdots, k+m) \qquad (1 \leqq k \leqq n-m-1).$$

For example, if the D_k all have the same sign and the d_k all have the same sign, then R has a band extension the absolute value of whose determinant is maximal over the class of all selfadjoint extensions of R that lie in the connected component of the band extension. This case includes the Toeplitz case. If the signs of D_k are alternating and the same is true of the signs of d_k, then the results are quite different. In particular, the absolute value of the determinant of the band extension is minimized rather than maximized. In the general case, the band extension is a stationary point for the determinant function, and can yield a maximum, a minimum, or a saddle point.

The paper is divided into five sections. In the first the situation for one-step extensions is analyzed. In the second we introduce the concepts of a band extension and a central extension and prove that they coincide. In §3 we investigate interior extensions, those in the connected component of the band extension. Another class of extensions, the so-called sign-consistent extensions, appears in § 4. The final section is dedicated to extensions of Toeplitz matrices.

1. One-step extensions. Let R be an $n \times n$ selfadjoint matrix with $R_{n1} = 0$. For any complex number w, let $F(w)$ be the $n \times n$ selfadjoint matrix such that $F(w)_{n1} = w$ and $F(w)_{jk} = R_{jk}$ for any pair of indices other than $(1, n)$ and $(n, 1)$. Then $F(w)$ is called a *one-step extension* of R. The two main problems discussed in this section are to determine the values of w for which $F(w)$ is invertible and to describe various properties of $\det F(w)$. We will assume that the determinants of $R(1, \cdots, n-1)$, $R(2, \cdots, n)$, and $R(2, \cdots, n-1)$ are not 0.

THEOREM 1.1. *Let R be an $n \times n$ selfadjoint matrix such that $R_{n1} = 0$ and the determinants of $R(1, \cdots, n-1)$, $R(2, \cdots, n)$, and $R(2, \cdots, n-1)$ are not* 0. *For any complex number w, let $F(w)$ be the selfadjoint extension of R such that $F(w)_{n1} = w$.*

(a) *If* $\det R(1, \cdots, n-1)$ *and* $\det R(2, \cdots, n)$ *have opposite signs, then $F(w)$ is invertible for all w. In that case,* $\det F(w)$ *and* $\det R(2, \cdots, n-1)$ *have opposite signs.*

(b) *If* $\det R(1, \cdots, n-1)$ *and* $\det R(2, \cdots, n)$ *have the same sign, then $F(w)$ is invertible for all w except those on a circle. The center w_0 and the radius ρ of the circle are determined as follows. Let*

$$P_{n-1} = \frac{\det R(1, \cdots, n-1)}{\det R(2, \cdots, n-1)}$$

and let $x_2, \cdots, x_{n-1}$ be the unique numbers satisfying

$$R(1, \cdots, n-1) \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} P_{n-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Then

$$w_0 = -\sum_{j=2}^{n-1} R_{nj} x_j \tag{1.1}$$

and

$$\rho = \frac{\sqrt{\det R(1, \cdots, n-1) \cdot \det R(2, \cdots, n)}}{|\det R(2, \cdots, n-1)|}. \tag{1.2}$$

Moreover, $\det F(w)$ *and* $\det R(2, \cdots, n-1)$ *have the same sign for* $|w - w_0| < \rho$, *but opposite signs for* $|w - w_0| > \rho$.

(c) *The entry* $[F(w)^{-1}]_{n1}$ *is* 0 *if and only if* $w = w_0$.

Proof. We have

$$F(w) = \begin{bmatrix} A & b^* \\ b & R_{nn} \end{bmatrix}$$

where $A = R(1, \cdots, n-1)$ and $b = [w\, R_{n2} \cdots R_{n,n-1}]$. Since A is invertible,

$$F(w) = \begin{bmatrix} I & 0 \\ bA^{-1} & 1 \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & v \end{bmatrix} \begin{bmatrix} I & A^{-1}b^* \\ 0 & 1 \end{bmatrix}$$

where $v = R_{nn} - bA^{-1}b^*$. Therefore, $\det F(w) = v \det A$, so that $F(w)$ is invertible if and only if $v \neq 0$, that is,

$$bA^{-1}b^* \neq R_{nn}. \tag{1.3}$$

Let

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{and} \quad b_1 = \begin{bmatrix} 0 \\ R_{2n} \\ \vdots \\ R_{n-1,n} \end{bmatrix}. \tag{1.4}$$

Then $b^* = \bar{w}e_1 + b_1$ and

$$\begin{aligned} bA^{-1}b^* &= \langle A^{-1}b^*, b^* \rangle = \alpha w\bar{w} + \beta w + \bar{\beta}\bar{w} + \gamma \\ &= \alpha \left| w + \frac{\bar{\beta}}{\alpha} \right|^2 + \gamma - \frac{|\beta|^2}{\alpha} \end{aligned}$$

where

$$\alpha = \langle A^{-1}e_1, e_1 \rangle = \frac{\det R(2, \cdots, n-1)}{\det R(1, \cdots, n-1)} \neq 0, \tag{1.5}$$

$$\beta = \langle A^{-1}b_1, e_1 \rangle, \qquad \gamma = \langle A^{-1}b_1, b_1 \rangle. \tag{1.6}$$

Therefore the condition in (1.3) for $F(w)$ to be invertible becomes

$$\left| w + \frac{\bar{\beta}}{\alpha} \right|^2 \neq \frac{R_{nn} - \gamma}{\alpha} + \frac{|\beta|^2}{\alpha^2}. \tag{1.7}$$

If the right side of (1.7) is negative, then $F(w)$ is invertible for all w. If the right side of (1.7) is nonnegative, then $F(w)$ is invertible for all w except those on the circle with center w_0 given by

$$w_0 = -\frac{\bar{\beta}}{\alpha} \tag{1.8}$$

and radius ρ equal to the square root of the right side of (1.7). Let

$$R(1, \cdots, n-1)^{-1} = (s_{jk})_{1 \leq j,k \leq n-1}.$$

Then from (1.4), (1.5), (1.6) and (1.8) it follows that

$$w_0 = -\frac{1}{s_{11}} \overline{\sum_{j=2}^{n-1} s_{1j}R_{jn}} = -\frac{1}{s_{11}} \sum_{j=2}^{n-1} R_{nj}s_{j1}. \tag{1.9}$$

Similarly, if $F(w)$ is partitioned into a 2×2 block matrix with $R(2, \cdots, n)$ in the lower right corner, and if we let

$$R(2, \cdots, n)^{-1} = (t_{jk})_{2 \leqq j,k \leqq n},$$

then

$$w_0 = -\frac{1}{t_{nn}} \overline{\sum_{j=2}^{n-1} R_{1j} t_{jn}}. \tag{1.10}$$

Now let

$$P_{n-1} = \frac{\det R(1, \cdots, n-1)}{\det R(2, \cdots, n-1)} \quad \text{and} \quad Q_{n-1} = \frac{\det R(2, \cdots, n)}{\det R(2, \cdots, n-1)} \tag{1.11}$$

so that

$$P_{n-1} = s_{11}^{-1} \quad \text{and} \quad Q_{n-1} = t_{nn}^{-1}. \tag{1.12}$$

Also, let $x_2, \cdots, x_{n-1}$ and $w_2, \cdots, w_{n-1}$ be the unique numbers satisfying

$$R(1, \cdots, n-1) \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} P_{n-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{1.13a}$$

and

$$R(2, \cdots, n) \begin{bmatrix} w_2 \\ \vdots \\ w_{n-1} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ Q_{n-1} \end{bmatrix} \tag{1.13b}$$

so that

$$x_j = P_{n-1} s_{j1} \quad \text{and} \quad w_j = Q_{n-1} t_{jn} \qquad (2 \leqq j \leqq n-1). \tag{1.14}$$

Let w be a complex number and define

$$\Delta_n' = w + \sum_{j=2}^{n-1} R_{nj} x_j \quad \text{and} \quad \Delta_n'' = \sum_{j=2}^{n-1} R_{1j} w_j + \bar{w}, \tag{1.15}$$

$$c_n' = -\frac{\Delta_n'}{Q_{n-1}} \quad \text{and} \quad c_n'' = -\frac{\Delta_n''}{P_{n-1}}, \tag{1.16}$$

$$P_n = P_{n-1} + c_n' \Delta_n'' \quad \text{and} \quad Q_n = Q_{n-1} + c_n'' \Delta_n'. \tag{1.17}$$

Then from (1.13)–(1.17) it follows that

$$F(w) \left(\begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{n-1} \\ 0 \end{bmatrix} + c_n' \begin{bmatrix} 0 \\ w_2 \\ \vdots \\ w_{n-1} \\ 1 \end{bmatrix} \right) = \begin{bmatrix} P_n \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{1.18}$$

and

$$(1.19) \qquad F(w)\left(c''_n\begin{bmatrix}1\\ x_2\\ \vdots\\ x_{n-1}\\ 0\end{bmatrix}+\begin{bmatrix}0\\ w_2\\ \vdots\\ w_{n-1}\\ 1\end{bmatrix}\right)=\begin{bmatrix}0\\ 0\\ \vdots\\ 0\\ Q_n\end{bmatrix}$$

so that

$$(1.20) \qquad P_n=\frac{\det F(w)}{\det R(2,\cdots,n)} \quad\text{and}\quad Q_n=\frac{\det F(w)}{\det R(1,\cdots,n-1)}.$$

Also (1.16) and (1.17) imply that

$$(1.21) \qquad P_n=P_{n-1}(1-c'_nc''_n) \quad\text{and}\quad Q_n=Q_{n-1}(1-c'_nc''_n).$$

It follows from (1.20) that $F(w)$ is invertible if and only if $P_n\neq 0$ or, equivalently, if and only if $Q_n\neq 0$. Therefore (1.21) implies that $F(w)$ is invertible if and only if

$$(1.22) \qquad c'_nc''_n\neq 1.$$

From (1.12) and (1.14) we find that (1.9) and (1.10) may be rewritten as

$$(1.23) \qquad w_0=-\sum_{j=2}^{n-1}R_{nj}x_j=-\overline{\sum_{j=2}^{n-1}R_{1j}w_j}$$

so that (1.15) becomes

$$(1.24) \qquad \Delta'_n=w-w_0 \quad\text{and}\quad \Delta''_n=\overline{w-w_0}.$$

From (1.16) and (1.24) it follows that

$$(1.25) \qquad |w-w_0|^2=c'_nc''_nP_{n-1}Q_{n-1}.$$

Therefore the condition in (1.22) for the invertibility of $F(w)$ becomes

$$(1.26) \qquad |w-w_0|^2\neq P_{n-1}Q_{n-1}=\frac{\det R(1,\cdots,n-1)\cdot\det R(2,\cdots,n)}{[\det R(2,\cdots,n-1)]^2}.$$

All statements in the theorem involving the invertibility of $F(w)$ follow from (1.23) and (1.26). For the statements involving $\det F(w)$, observe from (1.11), (1.20), (1.21) and (1.25) that

$$\det F(w)=\frac{\det R(1,\cdots,n-1)\cdot\det R(2,\cdots,n)}{\det R(2,\cdots,n-1)}$$
$$\times\left(1-\frac{|w-w_0|^2(\det R(2,\cdots,n-1))^2}{\det R(1,\cdots,n-1)\cdot\det R(2,\cdots,n)}\right).$$

We conclude that if $\det R(1,\cdots,n-1)$ and $\det R(2,\cdots,n)$ have opposite signs, then so do $\det F(w)$ and $\det R(2,\cdots,n-1)$, and that if $\det R(1,\cdots,n-1)$ and $\det R(2,\cdots,n)$ have the same sign, then

$$\det F(w)=\frac{\det R(1,\cdots,n-1)\cdot\det R(2,\cdots,n)}{\det R(2,\cdots,n-1)}\left(1-\frac{|w-w_0|^2}{\rho^2}\right)$$

so that $\det F(w)$ and $\det R(2,\cdots,n-1)$ have the same sign for $|w-w_0|<\rho$, but opposite signs for $|w-w_0|>\rho$. Finally, it follows from (1.18) that the first column of

$F(w)^{-1}$ is

$$P_n^{-1}\left(\begin{bmatrix}1\\x_2\\\vdots\\x_{n-1}\\0\end{bmatrix}+c_n'\begin{bmatrix}0\\w_2\\\vdots\\w_{n-1}\\1\end{bmatrix}\right).$$

Therefore $[F(w)^{-1}]_{n1}$ is 0 if and only if $c_n' = 0$, which means that $w = w_0$ by (1.16) and (1.24). This completes the proof of Theorem 1.1.

Suppose the matrix R in Theorem 1.1 is Toeplitz. Then

$$R(1, \cdots, n-1) = R(2, \cdots, n),$$

so it follows from (1.11) that $P_{n-1} = Q_{n-1}$ and from (1.16) and (1.24) that $c_n'' = \overline{c_n'}$. Then (1.21) becomes

$$P_n = P_{n-1}(1 - |c_n'|^2).$$

Thus the two sequences $\{c_n'\}$ and $\{c_n''\}$ may be regarded as a generalization of the single sequence of reflection coefficients that appear in the version of the Levinson algorithm in [3].

Let R be as in Theorem 1.1. The number w_0 given by (1.1) is called the *center of extension* of R. The *radius of extension* ρ of R is defined by (1.2) if det $R(1, \cdots, n-1)$ and det $R(2, \cdots, n)$ have the same sign and otherwise is defined to be ∞. The extension $F(w_0)$ is called the *central extension* of R, and $F(w)$ is said to be an *interior extension* if $|w - w_0| < \rho$. According to Theorem 1.1, every interior extension is invertible.

Statement (c) of Theorem 1.1 characterizes the central extension as being the unique invertible selfadjoint extension whose inverse has $(n, 1)$ entry equal to 0.

An alternate formula for the center of extension w_0 of R is given by

$$w_0 = [R_{n2} \cdots R_{n,n-1}]R(2, \cdots, n-1)^{-1}\begin{bmatrix}R_{21}\\\vdots\\R_{n-1,1}\end{bmatrix}.$$

To see this, we partition the minor of F_{1n} in $F(w_0)$ as follows:

$$\left[\begin{array}{c|ccc}R_{21} & R_{22} & \cdots & R_{2,n-1}\\\vdots & \vdots & & \vdots\\R_{n-1,1} & R_{n-1,2} & \cdots & R_{n-1,n-1}\\\hline w_0 & R_{n2} & \cdots & R_{n,n-1}\end{array}\right] = \begin{bmatrix}C & D\\w_0 & B\end{bmatrix}$$

where $D = R(2, \cdots, n-1)$. Then D is nonsingular, and

$$\begin{bmatrix}C & D\\w_0 & B\end{bmatrix} = \begin{bmatrix}I & 0\\BD^{-1} & 1\end{bmatrix}\begin{bmatrix}0 & D\\x & 0\end{bmatrix}\begin{bmatrix}1 & 0\\D^{-1}C & I\end{bmatrix} \tag{1.27}$$

where I is the $(n-2) \times (n-2)$ identity matrix, and $x = w_0 - BD^{-1}C$. If we consider the adjoint formula for $F(w_0)^{-1}$, it is clear from statement (c) of Theorem 1.1 that the matrix on the left of (1.27) is singular. This implies that $x = 0$, and hence $w_0 = BD^{-1}C$.

The next theorem characterizes the central extension by means of extremal properties of the absolute value of its determinant.

THEOREM 1.2. *Let R be an $n \times n$ selfadjoint matrix such that $R_{n1} = 0$ and the determinants of $R(1, \cdots, n-1)$, $R(2, \cdots, n)$, and $R(2, \cdots, n-1)$ are not* 0.

(a) *Suppose* det $R(1, \cdots, n-1)$ *and* det $R(2, \cdots, n)$ *have the same sign. Then for any interior extension $F(w)$ of R,*

$$|\det F(w)| \leqq |\det F(w_0)|$$

with equality only if $w = w_0$.

(b) *Suppose* det $R(1, \cdots, n-1)$ *and* det $R(2, \cdots, n)$ *have opposite signs. Then for any selfadjoint extension $F(w)$ of R,*

$$|\det F(w)| \geqq |\det F(w_0)|$$

with equality only if $w = w_0$.

(c) *If $F(w)$ is an interior extension of R, then* det $F(w)$ *and* det $F(w_0)$ *have the same sign.*

Proof. Observe from (1.11), (1.20) and (1.21) that

$$\det F(w) = \frac{\det R(1, \cdots, n-1) \cdot \det R(2, \cdots, n)}{\det R(2, \cdots, n-1)}(1 - c'_n c''_n) \tag{1.28}$$

so that

$$\det F(w) = \det F(w_0)(1 - c'_n c''_n). \tag{1.29}$$

If det $R(1, \cdots, n-1)$ and det $R(2, \cdots, n)$ have the same sign, then $0 < 1 - c'_n c''_n \leqq 1$, whereas if $\det R(1, \cdots, n-1)$ and $\det R(2, \cdots, n)$ have opposite signs, then $1 - c'_n c''_n \geqq 1$. Moreover, $1 - c'_n c''_n = 1$ only if $w = w_0$. The conclusions in (a) and (b) follow immediately from these remarks. Statement (c) is a simple corollary of Theorem 1.1.

The following result complements Theorem 1.1.

PROPOSITION 1.3. *Let R be an $n \times n$ selfadjoint matrix with $R_{n1} = 0$ and suppose that $R(1, \cdots, n-1)$ and $R(2, \cdots, n-1)$ are invertible with determinants of the same sign. Suppose* det $R(2, \cdots, n) = 0$. *If $F = F(w)$ is an invertible selfadjoint extension of R, then* det F *does not have the same sign as* det $R(1, \cdots, n-1)$.

Proof. We define P_{n-1} and $x_2, \cdots, x_{n-1}$ as before by (1.11) and (1.13). Since $R(2, \cdots, n)$ is singular but $R(2, \cdots, n-1)$ is nonsingular, the nullspace of $R(2, \cdots, n)$ is one-dimensional, so that there are unique $w_2, \cdots, w_{n-1}$ such that

$$R(2, \cdots, n)\begin{bmatrix} w_2 \\ \vdots \\ w_{n-1} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{1.30}$$

We now define Δ'_n, Δ''_n, and c''_n by (1.15) and (1.16), respectively. We also define

$$Q_n = c''_n \Delta'_n.$$

Then

$$Q_n = -\frac{\Delta''_n \Delta'_n}{P_{n-1}}. \tag{1.31}$$

We will now show that $\Delta''_n \Delta'_n \geqq 0$. For $\varepsilon > 0$, let

$$F_\varepsilon = F + \varepsilon I.$$

Since det $F(2, \cdots, n) = 0$ and det $F(1, \cdots, n-1)$ and det $F(2, \cdots, n-1)$ have the same sign, if $\varepsilon > 0$ is small, then det $F_\varepsilon(2, \cdots, n) \neq 0$ and det $F_\varepsilon(1, \cdots, n-1)$ and det $F_\varepsilon(2, \cdots, n-1)$ have the same sign. Let $x_2(\varepsilon), \cdots, x_{n-1}(\varepsilon)$, $P_{n-1}(\varepsilon)$, $w_2(\varepsilon), \cdots,$

$w_{n-1}(\varepsilon)$, $Q_{n-1}(\varepsilon)$, $\Delta'_n(\varepsilon)$, $\Delta''_n(\varepsilon)$, and $w_0(\varepsilon)$ be the numbers defined by (1.9), (1.11), (1.13) and (1.15) when R is replaced by $R_\varepsilon = R + \varepsilon I$. From (1.13) and (1.11) it follows that

$$\begin{bmatrix} w_2(\varepsilon) \\ \vdots \\ w_{n-1}(\varepsilon) \\ 1 \end{bmatrix} = \frac{\det R_\varepsilon(2, \cdots, n)}{\det R_\varepsilon(2, \cdots, n-1)} (\text{last column of } [R_\varepsilon(2, \cdots, n)]^{-1}).$$

But $\lim_{\varepsilon\to 0^+} \det R_\varepsilon(2, \cdots, n)$ (last column of $[R_\varepsilon(2, \cdots, n)]^{-1}$) exists, and

$$\lim_{\varepsilon\to 0^+} \det R_\varepsilon(2, \cdots, n-1) = \det R(2, \cdots, n-1) \neq 0.$$

Therefore there exist scalars $v_2, \cdots, v_{n-1}$ such that

$$\lim_{\varepsilon\to 0^+} \begin{bmatrix} w_2(\varepsilon) \\ \vdots \\ w_{n-1}(\varepsilon) \\ 1 \end{bmatrix} = \begin{bmatrix} v_2 \\ \vdots \\ v_{n-1} \\ 1 \end{bmatrix}.$$

Since $\lim_{\varepsilon\to 0^+} Q_{n-1} = 0$, it follows from (1.13) with R replaced by R_ε that

$$R(2, \cdots, n) \begin{bmatrix} v_2 \\ \vdots \\ v_{n-1} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

Therefore, (1.30) implies that

$$\begin{bmatrix} v_2 \\ \vdots \\ v_{n-1} \\ 1 \end{bmatrix} = \begin{bmatrix} w_2 \\ \vdots \\ w_{n-1} \\ 1 \end{bmatrix}$$

so that by (1.15)

$$\lim_{\varepsilon\to 0^+} \Delta''_n(\varepsilon) = \Delta''_n.$$

Similarly,

$$\lim_{\varepsilon\to 0^+} \Delta'_n(\varepsilon) = \Delta'_n.$$

Since $\Delta''_n(\varepsilon)\Delta'_n(\varepsilon) \geqq 0$ by (1.24), it follows that $\Delta''_n\Delta'_n \geqq 0$. Since $P_{n-1} > 0$ (because $\det R(1, \cdots, n-1)$ and $\det R(2, \cdots, n-1)$ have the same sign), it follows from (1.31) that $Q_n \leqq 0$. Moreover, it is easy to verify that (1.19) is satisfied by Q_n so that

$$\det F(w) = Q_n \det R(1, \cdots, n-1).$$

Therefore $\det F(w)$ does not have the same sign as $\det R(1, \cdots, n-1)$.

2. Band extensions, central extensions and the permanence principle. Let m and n be integers with $0 < m \leqq n - 2$. An $n \times n$ matrix R is called an *m-band matrix* if $R_{jk} = 0$ for $|k - j| > m$. Such a matrix will be called a *standard m-*band matrix if $R(j, \cdots, j + m)$ is invertible for $1 \leqq j \leqq n - m$ and $R(j + 1, \cdots, j + m)$ is invertible for $1 \leqq j \leqq n - m - 1$. In §1 we analyzed the situation for $(m + 2) \times (m + 2)$ selfadjoint standard m-band matrices. In this section we begin to generalize the results of §1 to $n \times n$ selfadjoint standard m-band matrices, for any $n \geqq m + 2$.

Let R be an $n \times n$ selfadjoint standard m-band matrix. An *extension* of R is an $n \times n$ matrix F such that $F_{jk} = R_{jk}$ for $|k - j| \leqq m$. A *band extension* of R is an invertible extension of R whose inverse is an m-band matrix. Let F be a selfadjoint extension of R. For $m + 2 \leqq i \leqq n$ and $1 \leqq j \leqq i - m - 1$, $F(j, \cdots, i)$ may be considered as a one-step extension of the $(i - j - 1)$-band matrix formed from $F(j, \cdots, i)$ by replacing F_{ij} and F_{ji} by 0. If for all such i and j, $F(j, \cdots, i)$ is an interior one-step extension, we say that F is an *interior extension* of R. If for all such i and j, $F(j, \cdots, i)$ is the central one-step extension, we call F the *central extension* of R.

THEOREM 2.1. *Let R be an $n \times n$ selfadjoint standard m-band matrix. Then the central extension F of R is the unique band extension of R. Moreover, F^{-1} admits unique factorizations of the form $F^{-1} = LM_1L^*$ and $F^{-1} = UM_2U^*$, where L (respectively, U) is a lower (respectively, upper) triangular m-band matrix with diagonal entries equal to 1 and with $L(n - m, \cdots, n)$ and $U(1, \cdots, m + 1)$ equal to the $(m + 1) \times (m + 1)$ identity matrix, and where*

$$M_1 = \begin{bmatrix} S & 0 \\ 0 & C \end{bmatrix} \quad \text{and} \quad M_2 = \begin{bmatrix} D & 0 \\ 0 & T \end{bmatrix}$$

with S and T $(n - m - 1) \times (n - m - 1)$ diagonal matrices, $C = R(n - m, \cdots, n)^{-1}$, and $D = R(1, \cdots, m + 1)^{-1}$. In fact,

$$S = \operatorname{diag}(P_j^{-1})_{1 \leqq j \leqq n-m-1} \quad \text{and} \quad T = \operatorname{diag}(Q_j^{-1})_{m+2 \leqq j \leqq n},$$

where

$$P_j = \frac{\det R(j, \cdots, j+m)}{\det R(j+1, \cdots, j+m)} \qquad (1 \leqq j \leqq n-m-1) \tag{2.1}$$

and

$$Q_j = \frac{\det R(j-m, \cdots, j)}{\det R(j-m, \cdots, j-1)} \qquad (m+2 \leqq j \leqq n) \tag{2.2}$$

and the entries of L in the band and below the diagonal are determined by the conditions $L(n - m, \cdots, n) = I$ and

$$R(j, \cdots, j+m) \begin{bmatrix} 1 \\ L_{j+1,j} \\ \vdots \\ L_{j+m,j} \end{bmatrix} = \begin{bmatrix} P_j \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (1 \leqq j \leqq n-m-1). \tag{2.3}$$

The entries of U in the band and above the diagonal are determined by the conditions $U(1, \cdots, m + 1) = I$ and

$$[\bar{U}_{j-m,j} \cdots \bar{U}_{j-1,j}\ 1] R(j, \cdots, j+m) = [0 \cdots 0\ Q_j] \qquad (m+2 \leqq j \leqq n). \tag{2.4}$$

Proof. We will first prove that the inverse of the central extension F is an m-band matrix. The proof is by induction on n. Suppose that $n = m + 2$. Using (1.17) and (1.18) and the fact that $c'_n = 0$ for a central one-step extension, we have

$$F \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{n-1} \\ 0 \end{bmatrix} = \begin{bmatrix} P_{n-1} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}. \tag{2.5}$$

From (2.5) it follows that F^{-1} is an m-band matrix. Now suppose $n > m + 2$ and assume the result is true for $(n-1) \times (n-1)$ matrices. We will first prove that the entries in the first column of F^{-1} that lie outside the band are 0. By the inductive hypothesis applied to $F(1, \cdots, n-1)$ and from the equation

$$R(1, \cdots, n-1) \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} P_{n-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

which appeared in (1.13), we have $x_k = 0$ for $m + 1 < k \leqq n - 1$, so (2.5) becomes

$$F \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{m+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} P_{n-1} \\ 0 \\ \\ \vdots \\ \\ 0 \end{bmatrix}$$

so that the entries in the first column of F^{-1} that lie outside the band are 0. Now let $D = R(2, \cdots, n)$ and let a be the $1 \times (n-1)$ matrix defined by

$$F\begin{bmatrix} 0 & 0 \\ 0 & D^{-1} \end{bmatrix} = \begin{bmatrix} 0 & a \\ 0 & I \end{bmatrix}.$$

Then

$$P_{n-1}^{-1} F \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{m+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} [1 \quad -a] = \begin{bmatrix} 1 & -a \\ 0 & 0 \end{bmatrix}.$$

Therefore

$$F\left(\begin{bmatrix} 0 & 0 \\ 0 & D^{-1} \end{bmatrix} + P_{n-1}^{-1} \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_{m+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} [1 \quad -a] \right) = I.$$

This implies that F^{-1} is an m-band matrix since D^{-1} is an m-band matrix. Therefore the inverse of the central extension is an m-band matrix, so the central extension is a band extension of R.

We next prove that the inverse of any band extension F of R has a unique factorization $F^{-1} = LM_1L^*$ of the form given in the statement of the theorem. For $\varepsilon > 0$ let

$$F_\varepsilon = F + \varepsilon\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$$

where I is the $(m+1) \times (m+1)$ identity matrix. For ε small and different from 0, F_ε is a standard m-band matrix and $F_\varepsilon(j, \cdots, n)$ is invertible for $n - m \leqq j \leqq n$. Thus by Theorem 3.1 of [5], F_ε^{-1} has a unique factorization of the form

$$F_\varepsilon^{-1} = L_\varepsilon M_\varepsilon L_\varepsilon^* \tag{2.6}$$

where L_ε is a lower triangular band matrix with diagonal entries equal to 1. From the construction of this factorization in [5], it follows that the first $n - m - 1$ columns of L_ε equal the corresponding columns of the lower triangular matrix L defined in the statement of the theorem. Furthermore, the construction in [5] shows that $M_\varepsilon(1, \cdots, n - m - 1) = S$, where S is as in the statement of the theorem. Thus (2.6) may be rewritten in the form

$$F_\varepsilon^{-1} = L \begin{bmatrix} S & 0 \\ 0 & C_\varepsilon \end{bmatrix} L^*$$

where C_ε is an $(m+1) \times (m+1)$ matrix. Since L is invertible and

$$\lim_{\varepsilon \to 0^+} F_\varepsilon^{-1} = \lim_{\varepsilon \to 0^+} (F + \varepsilon I)^{-1} = F^{-1}$$

it follows that $\lim_{\varepsilon \to 0^+} C_\varepsilon$ exists. Therefore F^{-1} has a factorization of the form $F^{-1} = LM_1L^*$ as given in the statement of the theorem. To prove the uniqueness of such a factorization, we suppose that $L_1M_1L_1^*$ and $L_2M'_1L_2^*$ are two such factorizations. Then

$$L_2^{-1}L_1 = M'_1L_2^*L_1^{*-1}M_1^{-1}.$$

Let $A = L_2^{-1}L_1$ and $B = M'_1L_2^*L_1^{*-1}M_1^{-1}$. Then $A(1, \cdots, n - m - 1)$ is lower triangular and $B(1, \cdots, n - m - 1)$ is upper triangular. Therefore $A(1, \cdots, n - m - 1)$ is diagonal and hence is the identity. Therefore $L_2 = L_1$ since $L_2(n - m, \cdots, n) = I = L_1(n - m, \cdots, n)$. It follows that $M'_1 = M_1$, so F^{-1} has a unique factorization of the form LM_1L^*. A similar proof shows that F^{-1} has a unique factorization of the form UM_2U^* as stated in the theorem.

Now we will prove that there is a unique band extension of R. Let F be a band extension of R. By what we just proved, F^{-1} has a factorization LML^* as in the statement of Theorem 2.1. Let

$$L^{-1} = \begin{bmatrix} A & 0 \\ B & I \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} S & 0 \\ 0 & C \end{bmatrix}.$$

Then

$$F = L^{*-1}M^{-1}L^{-1} = \begin{bmatrix} A^* & B^* \\ 0 & I \end{bmatrix} \begin{bmatrix} S^{-1} & 0 \\ 0 & C^{-1} \end{bmatrix} \begin{bmatrix} A & 0 \\ B & I \end{bmatrix}$$

$$= \begin{bmatrix} A^*S^{-1}A + B^*C^{-1}B & B^*C^{-1} \\ C^{-1}B & C^{-1} \end{bmatrix}.$$

It follows that $C = F(n - m, \cdots, n)^{-1} = R(n - m, \cdots, n)^{-1}$, so that C is unique. Now let $1 \leqq j \leqq n - m - 1$. From the equation

$$FL = L^{*-1}M^{-1}$$

it follows that

$$R(j, \cdots, j+m) \begin{bmatrix} 1 \\ L_{j+1,j} \\ \vdots \\ L_{j+m,j} \end{bmatrix} = \begin{bmatrix} S_{jj}^{-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (1 \leqq j \leqq n - m - 1).$$

Therefore

$$S_{ij}=\frac{\det R(j+1,\cdots,j+m)}{\det R(j,\cdots,j+m)} \qquad (1\leqq j\leqq n-m-1)$$

so that S is unique. Since each $R(j,\cdots,j+m)$ is invertible, L is also uniquely determined. Therefore there is a unique band extension of R. This completes the proof of Theorem 2.1.

We remark that the existence of a band extension of a standard m-band matrix was also established in [1].

From Theorem 2.1, we obtain the following generalization of the permanence principle [6, Thm. 4].

THEOREM 2.2. *Let F be the band extension of an $n \times n$ selfadjoint standard m-band matrix R. Then $F(j,\cdots,k)$ is the band extension of $R(j,\cdots,k)$ for $1\leqq j\leqq n-m-1$ and $j+m+1\leqq k\leqq n$.*

Proof. If "band extension" is replaced by "central extension," the result is true by definition. Thus the theorem follows immediately from Theorem 2.1.

3. Interior extensions and extremal properties of the determinant. Let R be an $n\times n$ selfadjoint standard m-band matrix and let F be a selfadjoint extension of R. If $F(j,\cdots,j+k)$ is invertible for $1\leqq j\leqq n-m-1$ and $m<k\leqq n-j$, then F is said to be *strongly invertible* (with respect to the band). Clearly every interior extension is strongly invertible. In the next theorem we prove that the interior extensions are precisely those extensions that can be connected to the central extension by a path in the set of all strongly invertible extensions of R.

THEOREM 3.1. *Let R be an $n\times n$ selfadjoint standard m-band matrix. Let $\mathfrak{S}$ be the set of all strongly invertible selfadjoint extensions of R. The connected component in $\mathfrak{S}$ of the central extension F_c of R is the set of all interior extensions of R.*

Proof. Let F be an interior extension of R. For $m+1\leqq k\leqq n$ let R_k be the $n\times n$ $(k-1)$-band matrix whose (i,j) entry is the (i,j) entry of F for $|i-j|\leqq k-1$. For $m+1\leqq k\leqq n-1$ and $0\leqq t\leqq 1$ let $R(k,t)$ be the $n\times n$ selfadjoint k-band matrix that agrees with R_k in the band of R_k and whose (i,j) entry for $i-j=k$ is $(1-t)F_{ij}+tC_{ij}$, where C_{ij} is the center of the one-step extension of $R_k(j,\cdots,i)$. By Theorem 1.1, $R(k,t)$ is a standard k-band matrix. For any $n\times n$ selfadjoint standard band matrix S let $C(S)$ be the central extension of S. From (1.1) it is clear that the center of a one-step extension is a continuous function of the other entries in the matrix. It follows that $C(S)$ is a continuous function of the entries in S, and $R(k,t)$ is a continuous function of t (for $0\leqq t\leqq 1$) and the entries in R_k. Therefore $C(R(k,t))$ is a continuous function of t for $0\leqq t\leqq 1$ that joins $C(R_{k+1})$ to $C(R_k)$ in $\mathfrak{S}$. Chaining together these functions for $k=n-1,\cdots,m+1$, we obtain a continuous function joining F to F_c in $\mathfrak{S}$. Therefore every interior extension is in the connected component of F_c.

Now let F be a strongly invertible selfadjoint extension of R that is not an interior extension. Then there are i and j such that $i-j>m$ and $F(j,\cdots,i)$ is not an interior one-step extension of $R_{i-j}(j,\cdots,i)$. By Theorem 1.1, det $F(j,\cdots,i)$ and det $C(R_{i-j}(j,\cdots,i))$ have opposite signs. Suppose there is a continuous function φ from $[0,1]$ into $\mathfrak{S}$ that joins F to F_c. Then $\varphi(t)(j,\cdots,i)$ is a continuous function of t that joins $F(j,\cdots,i)$ to the central extension of $R_{i-j}(j,\cdots,i)$ in the set of invertible $(i-j+1)\times(i-j+1)$ matrices. This contradicts the fact that det $F(j,\cdots,i)$ and det $C(R_{i-j}(j,\cdots,i))$ have opposite signs.

Let R be an $n\times n$ selfadjoint standard m-band matrix and let F be a strongly invertible selfadjoint extension of R. Let i and j be integers with $m+2\leqq i\leqq n$ and $1\leqq j\leqq i-m-1$, and consider the $(i-j-1)$-band matrix formed from $F(j,\cdots,i)$ by changing F_{ij} and F_{ji} to 0. If the radius of extension of this matrix is finite, we say that F has *finite radius* at (i, j) and (j, i). Otherwise F has *infinite radius* at (i, j) and (j, i).

By Theorem 1.1, F has finite radius at (i, j) and (j, i) if and only if

$$\det F(j, \cdots, i-1) \cdot \det F(j+1, \cdots, i) > 0.$$

THEOREM 3.2. *Let R be an $n \times n$ standard m-band matrix R. For any interior selfadjoint extension F of R, the positions at which F has infinite radius and the signs of the determinants*

$$\det F(j, \cdots, k) \qquad (1 \leqq j \leqq n-m-1, \quad j+m<k \leqq n)$$

are uniquely determined by the signs of the determinants

$$\det R(j, \cdots, j+m) \qquad (1 \leqq j \leqq n-m),$$

and

$$\det R(j+1, \cdots, j+m) \qquad (1 \leqq j \leqq n-m-1).$$

Proof. The proof is by induction on n. For $n = m + 2$ the result follows from Theorem 1.1. Suppose the result is true for $(n - 1) \times (n - 1)$ matrices. Then the positions in $F(1, \cdots, n-1)$ and $F(2, \cdots, n)$ at which F has infinite radius are determined by the signs of the given determinants in R. Also the signs of the determinants $\det F(j, \cdots, k)$ for $k > j + m$, and $k < n$ if $j = 1$, are determined. In particular, the signs of $\det F(1, \cdots, n-1)$, $\det F(2, \cdots, n)$, and $\det R(2, \cdots, n-1)$ are determined, and these, in turn, determine the sign of $\det F(1, \cdots, n)$ and whether F has infinite radius at $(1, n)$ and $(n, 1)$.

Let R be an $n \times n$ selfadjoint standard m-band matrix. We define

$$D_k = \det R(k, \cdots, k+m) \qquad (1 \leqq k \leqq n-m),$$

$$d_k = \det R(k+1, \cdots, k+m) \qquad (1 \leqq k \leqq n-m-1).$$

For any strongly invertible selfadjoint extension F of R and for $m + 2 \leqq i \leqq n$ and $1 \leqq j \leqq i - m - 1$ we let c'_{ij} and c''_{ij} be the numbers defined in (1.16) when $F(j, \cdots, i)$ is considered as a one-step extension.

THEOREM 3.3. *Let R be an $n \times n$ selfadjoint standard m-band matrix and let F be a selfadjoint extension of R. Then*

$$\det F = \frac{D_1 \cdots D_{n-m}}{d_1 \cdots d_{n-m-1}} \prod (1 - c'_{ij} c''_{ij}) \tag{3.1}$$

where the product is taken over all i, j such that $m + 2 \leqq i \leqq n$ and $1 \leqq j \leqq i - m - 1$.

Proof. The proof is by induction on n. For $n = m + 2$, it follows from (1.28) that

$$\det F = \frac{\det R(1, \cdots, n-1) \cdot \det R(2, \cdots, n)}{\det R(2, \cdots, n-1)} (1 - c'_{n1} c''_{n1})$$

$$= \frac{D_1 D_2}{d_1} (1 - c'_{n1} c''_{n1}).$$

Let $n > m + 2$ and assume the result is true for $k \times k$ matrices with $m + 2 \leqq k \leqq n - 1$. Then

$$\det F(1, \cdots, n-1) = \frac{D_1 \cdots D_{n-m-1}}{d_1 \cdots d_{n-m-2}} \prod (1 - c'_{ij} c''_{ij})$$

where the product is over all i, j such that $m + 2 \leqq i \leqq n - 1$ and $1 \leqq j \leqq i - m - 1$;

$$\det F(2, \cdots, n) = \frac{D_2 \cdots D_{n-m}}{d_2 \cdots d_{n-m-1}} \prod (1 - c'_{ij} c''_{ij})$$

where the product is over all i, j such that $m + 3 \leqq i \leqq n$ and $2 \leqq j \leqq i - m - 1$; and

$$\det F(2, \cdots, n-1) = \frac{D_2 \cdots D_{n-m-1}}{d_2 \cdots d_{n-m-2}} \prod (1 - c'_{ij} c''_{ij})$$

where the product is over all i, j such that $m + 3 \leqq i \leqq n - 1$ and $2 \leqq j \leqq i - m - 1$. From these formulas it follows that

$$\det F = \frac{\det F(1, \cdots, n-1) \cdot \det F(2, \cdots, n)}{\det F(2, \cdots, n-1)} (1 - c'_{n1} c''_{n1})$$

$$= \frac{D_1 \cdots D_{n-m}}{d_1 \cdots d_{n-m-1}} \prod (1 - c'_{ij} c''_{ij})$$

where the product is over all i, j such that $m + 2 \leqq i \leqq n$ and $1 \leqq j \leqq i - m - 1$.

The determinant formula (3.1) generalizes a formula in [1]. See also [2].

The next theorem is a generalization of the maximum entropy principle [3], [5], [6]. Recall that for $m + 2 \leqq i \leqq n$ and $1 \leqq j \leqq i - m - 1$, $F(j, \cdots, i)$ may be viewed as a one-step extension.

THEOREM 3.4. *Let R be an $n \times n$ selfadjoint standard m-band matrix. Let F be an interior extension of R with the property that if F has infinite radius at (i, j), then F_{ij} is the center of the corresponding one-step extension. Then*

$$|\det F_c| \geqq |\det F|$$

with equality only if $F = F_c$.

Proof. We use (3.1) along with the fact that $1 - c'_{ij} c''_{ij} > 0$ for interior extensions, $1 - c'_{ij} c''_{ij} = 1$ if F_{ij} is the corresponding center, and $1 - c'_{ij} c''_{ij} \leqq 1$ if (i, j) has finite radius with equality only if F_{ij} is the corresponding center. It follows that

$$|\det F| \leqq \left| \frac{D_1 \cdots D_{n-m}}{d_1 \cdots d_{n-m-1}} \right| = |\det F_c|$$

with equality only if $F = F_c$.

From the determinant formula (3.1) we also obtain the following minimum entropy principle.

THEOREM 3.5. *Let R be an $n \times n$ selfadjoint standard m-band matrix. Let F be an interior extension of R with the property that if F has finite radius at (i, j), then F_{ij} is the corresponding center. Then*

$$|\det F_c| \leqq |\det F|$$

with equality only if $F = F_c$.

Proof. The result follows from (3.1) and the fact that if F has infinite radius at (i, j), then $1 - c'_{ij} c''_{ij} \geqq 1$ with equality only if F_{ij} is the corresponding center.

4. Sign-consistent extensions.

There is interest in the inertia of extensions of selfadjoint band matrices [4], [7], [8]. In this section we investigate a class of such extensions whose inertia are determined by the given band matrix.

THEOREM 4.1. *Let R be an $n \times n$ selfadjoint standard m-band matrix, and let F be an interior extension of R. Then*

$$\operatorname{sgn} \frac{\det F(1, \cdots, j)}{\det F(1, \cdots, j-1)} = \operatorname{sgn} \frac{\det R(j-m, \cdots, j)}{\det R(j-m, \cdots, j-1)} \qquad (m+2 \leqq j \leqq n).$$

Proof. Let F_c be the central extension of R. Since F_c is the band extension of R,

F_c^{-1} has the factorization

$$F_c^{-1}=U\begin{bmatrix} R(1,\cdots,m+1)^{-1} & 0 \\ 0 & T \end{bmatrix}U^*$$

where U is an upper-triangular band matrix with diagonal entries equal to 1 and $T =$ diag $(Q_j^{-1})_{m+2\leqq j\leqq n}$, where

$$Q_j=\frac{\det R(j-m,\cdots,j)}{\det R(j-m,\cdots,j-1)} \qquad (m+2\leqq j\leqq n).$$

Then

$$F_c=U^{-1*}\begin{bmatrix} R(1,\cdots,m+1) & 0 \\ 0 & T^{-1} \end{bmatrix}U^{-1}.$$

Since U^{-1*} is lower-triangular it follows that

$$F_c(1,\cdots,j)=U^{-1*}(1,\cdots,j)\begin{bmatrix} R(1,\cdots,m+1) & 0 \\ 0 & T^{-1}(m+2,\cdots,j) \end{bmatrix}U^{-1}(1,\cdots,j)$$

for $m + 2 \leqq j \leqq n$, so that

$$\frac{\det F_c(1,\cdots,j)}{\det F_c(1,\cdots,j-1)}=Q_j=\frac{\det R(j-m,\cdots,j)}{\det R(j-m,\cdots,j-1)} \qquad (m+2\leqq j\leqq n).$$

Now let F be any interior extension. By Theorem 3.1, F lies in the same connected component of $\mathfrak{S}$ as F_c. It follows that for any k, $F(1, \cdots, k)$ is in the same connected component as $F_c(1, \cdots, k)$, so that

$$\operatorname{sgn}\frac{\det F(1,\cdots,j)}{\det F(1,\cdots,j-1)}=\operatorname{sgn}\frac{\det F_c(1,\cdots,j)}{\det F_c(1,\cdots,j-1)} \qquad (m+2\leqq j\leqq n).$$

This proves the theorem.

In view of Theorem 4.1, we make the following definition. Let R be an $n \times n$ selfadjoint standard m-band matrix and F a selfadjoint extension of F such that $F(1, \cdots, j)$ is invertible for $m + 2 \leqq j \leqq n$. Then F is *sign-consistent* if

$$\operatorname{sgn}\frac{\det F(1,\cdots,j)}{\det F(1,\cdots,j-1)}=\operatorname{sgn}\frac{\det R(j-m,\cdots,j)}{\det R(j-m,\cdots,j-1)} \qquad (m+2\leqq j\leqq n).$$

Theorem 4.1 states that every interior extension of a selfadjoint standard m-band matrix is sign-consistent.

To simplify notation we define the pattern matrix $P_F = (p_{jk})$ of an $n \times n$ matrix F by

$$p_{jk}=p_{kj}=\operatorname{sgn}\det F(j,\cdots,k) \qquad (1\leqq j\leqq k\leqq n).$$

If F is a strongly invertible selfadjoint extension of an $n \times n$ standard m-band matrix R, then for $k - j > m$ the entries p_{jk}, $p_{j,k-1}$, $p_{j+1,k}$ and $p_{j+1,k-1}$ in the pattern matrix P_F determine whether the matrix $F(j, \cdots, k)$ is an interior one-step extension. In fact, the following criteria follow immediately from Theorem 1.1:

(1°) If $p_{j,k-1}p_{j+1,k} = -1$, then $p_{jk}p_{j+1,k-1} = -1$ and $F(j, \cdots, k)$ is an interior one-step extension with infinite radius.

(2°) If $p_{j,k-1} = p_{j+1,k-1} = p_{j,k}$, then all three equal $p_{j+1,k}$ and $F(j, \cdots, k)$ is an interior one-step extension with finite radius.

(3°) The matrix $F(j, \cdots, k)$ is an interior one-step extension if and only if $p_{j,k-1}p_{j+1,k}p_{j,k}p_{j+1,k-1} = 1$.

The next two theorems give sufficient conditions on a standard m-band matrix R for every sign-consistent strongly invertible selfadjoint extension of R to be an interior extension of R.

THEOREM 4.2. *Let R be an $n \times n$ selfadjoint standard m-band matrix such that D_k/d_k is positive for $1 \leqq k \leqq n - m - 1$. Then a strongly invertible extension F of R is sign-consistent if and only if F is an interior extension. In that case, for each k with $m + 2 \leqq k \leqq n$, the entries p_{jk} in P_F have the same sign for $1 \leqq j \leqq k - m$.*

Proof. By Theorem 4.1 every interior extension is sign-consistent. Now suppose that F is sign-consistent and observe that

$$p_{k,k+m} = \operatorname{sgn} D_k \qquad (1 \leqq k \leqq n-m)$$

and

$$p_{k+1,k+m} = \operatorname{sgn} d_k \qquad (1 \leqq k \leqq n-m-1).$$

The proof is by induction on n. Suppose first that $n = m + 2$. Since F is sign-consistent, we have

$$\frac{p_{1,n}}{p_{1,n-1}} = \frac{p_{2,n}}{p_{2,n-1}} \tag{4.1}$$

and since $D_1/d_1 > 0$, we have $p_{1,n-1} = p_{2,n-1}$. Therefore, $p_{1,n} = p_{2,n}$, and it follows from (3°) that F is an interior extension. Now assume the result for $(n - 1) \times (n - 1)$ matrices. Then $F(1, \cdots, n - 1)$ is a strongly invertible sign-consistent extension of $R(1, \cdots, n - 1)$ and $P_F(1, \cdots, n - 1)$, which equals $P_{F(1, \cdots, n-1)}$, has the desired form. In particular, the entries $p_{j,n-1}$ are equal for $1 \leqq j \leqq n - m - 1$. There are two cases to consider: either $p_{1,n} = p_{1,n-1}$ or $p_{1,n} = -p_{1,n-1}$. If $p_{1,n} = p_{1,n-1}$, then repeated applications of criterion (2°) imply that the entries $p_{j,n}$ are equal for $1 \leqq j \leqq n - m$ and that F is an interior extension. If $p_{1,n} = -p_{1,n-1}$, then repeated applications of criterion (3°) imply the same conclusion. This completes the proof.

THEOREM 4.3. *Let R be an $n \times n$ selfadjoint standard m-band matrix such that D_k/d_k is negative for $1 \leqq k \leqq n - m - 1$. Then a strongly invertible extension F of R is sign-consistent if and only if F is an interior extension. In that case, for each k with $m + 1 \leqq k \leqq n$ the entries $p_{1,k}, \cdots, p_{k-m,k}$ form an alternating sequence.*

Proof. By Theorem 4.1 it suffices to prove the necessity. Suppose that F is sign-consistent. The proof is by induction on n. Suppose that $n = m + 2$. Then $p_{1,n-1} = -p_{2,n-1}$ and (4.1) holds. Therefore $p_{1,n} = -p_{2,n}$, and along with (3°) this implies that F is an interior extension. Now assume the result is true for $(n - 1) \times (n - 1)$ matrices. Then as in the proof of Theorem 4.2 we find that $F(1, \cdots, n - 1)$ is an interior extension of $R(1, \cdots, n - 1)$ and that the entries in $P_F(1, \cdots, n - 1)$ have the desired pattern. In particular, the entries $p_{1,n-1}, p_{2,n-1}, \cdots, p_{n-m-1,n-1}$ alternate in sign. If $p_{1,n} = -p_{1,n-1}$, then repeated applications of (1°) imply that F is an interior extension of R with $p_{1,n}, \cdots, p_{n-m,n}$ alternating in sign. If $p_{1,n} = p_{1,n-1}$, then $p_{n-m,n} = p_{n-m,n-1}$ since F is an interior extension of R with $p_{1,n}, \cdots, p_{n-m,n}$ alternating in sign. This completes the proof.

Theorems 4.2 and 4.3 do not cover every standard m-band matrix with the property that each of its strongly invertible sign-consistent extensions is interior. For example, let R be the following standard 1-band matrix:

$$\begin{bmatrix} 1 & 2 & 0 & 0 & 0 \\ 2 & 1 & 2 & 0 & 0 \\ 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 2 & 1 \end{bmatrix}.$$

It is easily seen from Theorem 1.1 that every strongly invertible sign-consistent extension of R is interior. However, D_1/d_1 and D_2/d_2 are negative, whereas D_3/d_3 is positive, so that R is not covered by either Theorem 4.2 or Theorem 4.3.

Furthermore, not every strongly invertible sign-consistent extension of a standard m-band matrix is interior. For example, if

$$R=\begin{bmatrix}-2 & -1 & 0 & 0\\ -1 & -1 & 2 & 0\\ 0 & 2 & -1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \quad\text{and}\quad F=\begin{bmatrix}-2 & -1 & 1 & 10\\ -1 & -1 & 2 & 3\\ 1 & 2 & -1 & 0\\ 10 & 3 & 0 & 1\end{bmatrix}$$

then R is a standard 1-band matrix and F is a strongly invertible sign-consistent extension of R. However, F is not an interior extension of R. Since the interior extensions form a connected component in the set $\mathfrak{S}$ of strongly invertible extensions (by Theorem 3.1), this example shows that the set of sign-consistent extensions may have more than one connected component.

We conclude this section by considering special cases of Theorems 4.2 and 4.3 along with corresponding entropy results for the central extension F_c.

THEOREM 4.4. *Let R be an $n \times n$ selfadjoint standard m-band matrix such that $D_1, \cdots, D_{n-m}$ have the same sign and $d_1, \cdots, d_{n-m-1}$ have the same sign. Then a strongly invertible selfadjoint extension F of R is sign-consistent if and only if F is an interior extension with finite radius at every position outside the band. Moreover,*

$$|\det F| \leqq |\det F_c|$$

with equality only if $F = F_c$.

Proof. By Theorem 4.1 every interior extension of R is sign-consistent. Let F be a strongly invertible sign-consistent selfadjoint extension of R. Since $D_1, \cdots, D_{n-m}$ have the same sign and $d_1, \cdots, d_{n-m-1}$ have the same sign, it follows that the hypotheses of either Theorem 4.2 or 4.3 are satisfied, and hence that F is an interior extension of R. If $D_k/d_k > 0$ for $1 \leqq k \leqq n-m-1$, then it is easy to see from the proof of Theorem 4.2 that the entries p_{jk} of P_F are the same for $|k-j| > m$, which implies that F has finite radius at every position outside the band. On the other hand, if $D_k/d_k < 0$ for $1 \leqq k \leqq n-m-1$, the proof of Theorem 4.3 shows that the entries p_{jk} with $|k-j| > m$ form a checkerboard pattern, which also implies that F has finite radius at every position outside the band. Thus in either case it follows from Theorem 3.4 that

$$|\det F| \leqq |\det F_c|$$

with equality only if $F = F_c$.

THEOREM 4.5. *Let R be an $n \times n$ selfadjoint standard m-band matrix such that $D_1, \cdots, D_{n-m}$ and $d_1, \cdots, d_{n-m-1}$ are each alternating sequences. Then a strongly invertible selfadjoint extension F of R is sign-consistent if and only if F is an interior extension with infinite radius at every position outside the band. Moreover,*

$$|\det F| \geqq |\det F_c|$$

with equality only if $F = F_c$.

Proof. As in the proof of Theorem 4.4 it follows that F is a strongly invertible sign-consistent extension of R if and only if F is an interior extension. If $D_k/d_k > 0$ for $1 \leqq k \leqq n-m-1$, then the proof of Theorem 4.2 shows that the entries p_{jk} above the band are constant in each column, with entries in adjacent columns having opposite signs. This implies that F has infinite radius at every position outside the band. If $D_k/d_k < 0$ for $1 \leqq k \leqq n-m-1$, then the proof of Theorem 4.3 shows that the entries p_{jk} above the band are constant in each row, with entries in adjacent rows having opposite signs.

This also implies that F has infinite radius at every position outside the band. Thus in either case it follows from Theorem 3.5 that

$$|\det F| \geqq |\det F_c|$$

with equality only if $F = F_c$.

5. Extensions of Toeplitz band matrices. Let R be an $n \times n$ selfadjoint Toeplitz standard m-band matrix. It is not difficult to prove by induction that the central extension of R is also Toeplitz. Since $R(j, \cdots, j+m) = R(1, \cdots, m+1)$ for $1 \leqq j \leqq n-m$ and $R(j+1, \cdots, j+m) = R(2, \cdots, m+1)$ for $1 \leqq j \leqq n-m-1$, we need to consider only two possibilities:

(i) $\det R(1, \cdots, m+1)$ and $\det R(2, \cdots, m+1)$ have the same sign;

(ii) $\det R(1, \cdots, m+1)$ and $\det R(2, \cdots, m+1)$ have opposite signs.

In either case it follows from Theorem 4.4 that every sign-consistent extension is an interior extension with finite radius at every position outside the band, since it is clear that a sign-consistent selfadjoint extension of a Toeplitz matrix is strongly invertible. Furthermore, in view of Theorem 4.1, this implies that a selfadjoint extension of R is sign-consistent if and only if it is interior. In case (i), if F is an interior extension of R, then $\det F(j, \cdots, k)$ has the same sign as $\det R(1, \cdots, m+1)$ for $1 \leqq j \leqq n-m-1$ and $j+m+1 \leqq k \leqq n$. In case (ii), if F is an interior extension of R, then for $1 \leqq j \leqq n-m-1$ and $j+m+1 \leqq k \leqq n$, $\det F(j, \cdots, k)$ has the same sign as $\det R(1, \cdots, m+1)$ if $k-j-m$ is even, and has the same sign as $\det R(2, \cdots, m+1)$ if $k-j-m$ is odd. In both cases, the absolute value of the determinant of the band extension is greater than the absolute value of the determinant of any other interior extension of R. These results are summarized in the following theorem.

THEOREM 5.1. *Let R be an $n \times n$ selfadjoint Toeplitz standard m-band matrix.*

(a) *The band extension F of R is Toeplitz.*

(b) *A selfadjoint extension of R is sign-consistent if and only if it is interior.*

(c) *Every sign-consistent extension of R has finite radius at every position outside the band.*

(d) *For any sign-consistent extension G of R,*

$$|\det G| \leqq |\det F|$$

with equality only if $G = F$.

REFERENCES

[1] W. W. BARRETT AND P. J. FEINSILVER, *Inverses of banded matrices*, Linear Algebra Appl., 41 (1981), pp. 111–130.

[2] W. W. BARRETT AND C. R. JOHNSON, *Determinantal formulas for matrices with sparse inverses*, Linear Algebra Appl., 56 (1984), pp. 73–88.

[3] J. BURG, *Maximum entropy spectral analysis*, Ph.D. dissertation, Stanford University, Stanford, CA, 1975.

[4] J. DANCIS, *The possible inertias for an Hermitian matrix and its principal submatrices*, Linear Algebra Appl., 85 (1987), pp. 121–151.

[5] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[6] R. L. ELLIS, I. GOHBERG AND D. LAY, *Band extensions, maximum entropy and the permanence principle*, in Maximum Entropy and Bayesian Methods in Applied Statistics, J. Justice, ed., Cambridge University Press, Cambridge, 1986.

[7] R. GRONE, C. R. JOHNSON, E. M. SÁ AND H. WOLKOWICZ, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[8] C. R. JOHNSON AND L. RODMAN, *Inertial possibilities for completions of partial Hermitian matrices*, Linear and Multilinear Algebra, 16 (1984), pp. 179–195.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

016

A PROJECTION DECOMPOSITION FOR BIVARIATE DISCRETE PROBABILITY DISTRIBUTIONS*

DEVENDRA CHHETRY† AND ALLAN R. SAMPSON†‡

Abstract. Let $Q = \{\text{Prob}\,(X = x_i, Y = y_j)\}$ for $x_1 < \cdots < x_m, y_1 < \cdots < y_n$. A new matrix decomposition of Q is given in terms of certain projections on linear spaces related to the marginal probabilities. It is shown that this decomposition implies Fisher's canonical decomposition and also a representation important in positive dependence. Also considered are applications of these ideas to the concordant monotone correlation, the maximal correlation and Hotelling's canonical correlation.

Key words. probability decomposition, bivariate ordinal random variables, maximal correlation, canonical correlation, concordant monotone correlation, positive quadrant dependence

AMS(MOS) subject classifications. Primary 15A51; secondary 62H20

1. Introduction and motivation. Let Q denote the probability mass function matrix of two random variables X and Y on the lattice $(x_1, \cdots, x_m) \times (y_1, \cdots, y_n)$, i.e., $Q = \{\text{Prob}\,(X = x_i, Y = y_j)\}$. The canonical decomposition of Q in terms of its X and Y marginal distributions (Fisher (1940), Maung (1941) and Lancaster (1958)) plays an important role in the structural analysis of bivariate distributions and in the analysis of contingency tables. For instance, Lancaster (1969) discusses the structural interpretation of the canonical decomposition, and Gilula (1984) indicates its application to the analysis of contingency tables. An important property of the canonical decomposition is that the corresponding second canonical correlation is the maximal correlation, a measure of association introduced by Hirschfeld (1935), and the second pair of canonical variables are the optimal scales for nominal contingency tables. This property is fundamental to the development of correspondence analysis (Benzecri (1973)) and dual scaling (Nishisato (1980)). More recently, motivated by ordinal contingency tables, research has focused on the study of monotonic relationships between X and Y (see Kimeldorf and Sampson (1978), Kimeldorf, May and Sampson (1982) and Nishisato and Arri (1975)). More generally, there is extensive literature concerning positive dependence among random variables, e.g., Tong (1980). With the exception of the recent work of Schriever (1985), the canonical decomposition apparently has not been used to study monotonic relationships and positive dependence. The purpose of this note is to obtain a new and more general decomposition of Q and to show that special cases of this decomposition yield the canonical decomposition, as well as a decomposition implicit in positive dependence.

Throughout we use the following notation. The row totals (X-marginal p.m.f.) of Q are denoted by the vector $\mathbf{r} = (r_1, \cdots, r_m)'$ and the column marginals by $\mathbf{c} = (c_1, \cdots, c_n)'$. For meaningfulness it is assumed that $r_i > 0$, $i = 1, \cdots, m$ and that $c_j > 0$, $j = 1, \cdots, n$; this is denoted by $\mathbf{r} > 0$ and $\mathbf{c} > 0$. For a vector $\mathbf{x} = (x_1, \cdots, x_p)'$, $\mathbf{x}^{1/2}$ denotes $(x_1^{1/2}, \cdots, x_p^{1/2})'$, and D_x denotes the diagonal matrix Diag $(x_1, \cdots, x_p)$. The p-dimensional vector $(1, \cdots, 1)'$ is denoted by $\mathbf{1}_p$ or simply $\mathbf{1}$; $J_p \equiv \mathbf{1}_p\mathbf{1}'_p$; and $\mathbf{e}_k$ denotes the kth coordinate unit vector. If V_1 and V_2 are two vector spaces such that $V_1 \oplus V_2 = \mathbb{R}^p$, then $P_{V_1}^{V_2}$ denotes the projection matrix onto V_1 along the direction of

* Received by the editors June 30, 1986; accepted for publication February 4, 1987. This work was supported by the Air Force Office of Scientific Research under contract 84-0113.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

‡ The research of this author was done in part at the Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

V_2. If $V_2 = V_1^\perp$, then we use P_{V_1} for $P_{V_1}^{V_1^\perp}$ where $V_1^\perp$ is the orthocomplement of V_1. Observe that $(P_{V_1}^{V_2})' = P_{V_2^\perp}^{V_1^\perp}$ and $(I - P_{V_1}^{V_2}) = P_{V_2}^{V_1}$. Finally, for a matrix A, $\mathscr{L}(A)$ denotes the linear space spanned by the columns of A.

In terms of Q, the canonical decomposition can be expressed as follows where it is assumed $m \leqq n$. Let $Q^* \equiv D_r^{-1/2} Q D_c^{-1/2}$ and write Q^* in terms of its spectral decomposition as

$$Q^* = \Gamma[\operatorname{Diag}(1, \rho_1, \cdots, \rho_{m-1}) : 0_{m,n-m}]G' \tag{1.1}$$

where $\Gamma = [D_r^{1/2}\mathbf{1}_m : \Gamma_1]$ and $G = [D_c^{1/2}\mathbf{1}_n : G_1]$ are orthogonal matrices, $0_{m,n-m}$ is an $m \times (n - m)$ matrix of zeros, and $1 \geqq \rho_1^2 \geqq \cdots \geqq \rho_{m-1}^2 \geqq 0$ are the eigenvalues of $Q^{*\prime}Q^*$ (or $Q^*Q^{*\prime}$). Then the canonical decomposition of Q can be derived from (1.1) as

$$Q = \mathbf{r}\mathbf{c}' + D_r^{1/2}\Gamma_1 D_\rho (D_c^{1/2} G_1)' \tag{1.2}$$

where $D_\rho = [\operatorname{Diag}(\rho_1, \cdots, \rho_{m-1}) : 0_{m-1,n-m}]$, and $1 \geqq \rho_1 \geqq \cdots \geqq \rho_{m-1} \geqq 0$ are the canonical correlations of Q. Sarmanov (1958a), (1958b) also calls ρ_1 the maximal correlation.

In the study of monotone dependence and positive (negative) quadrant dependence (Lehmann (1966)), in particular, the quantities Prob $(X > x, Y > y)$ − Prob $(X > x)$ × Prob $(Y > y)$ play an important role. It is straightforward to show that Q admits a decomposition in terms of these quantities, namely,

$$Q = \mathbf{r}\mathbf{c}' + \Delta_m(\bar{H} - \bar{F}\bar{G}')\Delta_n' \tag{1.3}$$

where the $(m - 1) \times (n - 1)$ matrix $\bar{H} = \{\sum_{k>i} \sum_{l>j} q_{kl}\}$, the $(m - 1)$-dimensional vector $\bar{\mathbf{F}} = (\bar{F}_i) \equiv (\sum_{k>i} r_k)$, and the $(n - 1)$-dimensional vector $\bar{\mathbf{G}} = (\bar{G}_i) \equiv (\sum_{j>i} c_j)$, and

$$\Delta_p = (\mathbf{e}_2 - \mathbf{e}_1 :, \cdots, : \mathbf{e}_p - \mathbf{e}_{p-1}). \tag{1.4}$$

Note that equivalently $\bar{H} - \bar{F}\bar{G}'$ can be expressed as $\{\sum_{k \leqq i} \sum_{l \leqq j} q_{kl} - \sum_{k \leqq i} r_k \sum_{l \leqq j} c_l\}$. We term this interesting decomposition given by (1.3) the quadrant dependence decomposition.

2. The general projection decomposition. A straightforward theorem for decomposing Q is proved and the general properties of this decomposition are studied. In § 3, it is shown that this decomposition yields as special cases the canonical decomposition of (1.2) and the quadrant dependence decomposition of (1.3).

THEOREM 2.1 (Projection Decomposition). *Let Q be an $m \times n$ probability matrix with $\mathbf{r} > 0$ and $\mathbf{c} > 0$. Then*

$$Q = P_{\mathscr{L}(\mathbf{r})}^{\mathscr{L}^\perp(\mathbf{1})} Q P_{\mathscr{L}(\mathbf{1})}^{\mathscr{L}^\perp(\mathbf{c})} + (I_m - P_{\mathscr{L}(\mathbf{r})}^{\mathscr{L}^\perp(\mathbf{1})}) Q (I_n - P_{\mathscr{L}(\mathbf{1})}^{\mathscr{L}^\perp(\mathbf{c})}). \tag{2.1}$$

Proof. Let $\mathbf{Q}_i$, $i = 1, \cdots, n$ denote the columns of Q. Then, for $i = 1, \cdots, n$

$$\mathbf{Q}_i = (c_i\mathbf{r}) + \left((1 - c_i)\mathbf{r} - \sum_{j \neq i} \mathbf{Q}_j\right).$$

Because $((1 - c_i)\mathbf{r} - \sum_{j \neq i} \mathbf{Q}_j)'\mathbf{1} = 0$ and $c_i\mathbf{r} \in \mathscr{L}(\mathbf{r})$, it follows that

$$P_{\mathscr{L}(\mathbf{r})}^{\mathscr{L}^\perp(\mathbf{1})} Q = \mathbf{r}\mathbf{c}'.$$

Now use the analogous result for $P_{\mathscr{L}(\mathbf{c})}^{\mathscr{L}^\perp(\mathbf{1})}$ to obtain

$$Q P_{\mathscr{L}(\mathbf{1})}^{\mathscr{L}^\perp(\mathbf{c})} = ((P_{\mathscr{L}(\mathbf{1})}^{\mathscr{L}^\perp(\mathbf{c})})' Q')' = (P_{\mathscr{L}(\mathbf{c})}^{\mathscr{L}^\perp(\mathbf{1})} Q')' = (\mathbf{c}\mathbf{r}')' = \mathbf{r}\mathbf{c}'.$$

Finally note $P_{\mathscr{L}(\mathbf{r})}^{\mathscr{L}^\perp(\mathbf{1})}(\mathbf{r}\mathbf{c}') = (\mathbf{r}\mathbf{c}')$, so that the result follows from simple algebra.

Remark 2.2. An alternate expression for (2.1) is

$$Q=\mathbf{rc}'+(P^{\mathscr{L}(\mathbf{r})}_{\mathscr{L}^{\perp}(\mathbf{1})})Q(P^{\mathscr{L}(\mathbf{1})}_{\mathscr{L}^{\perp}(\mathbf{c})}). \tag{2.2}$$

This expression relates a probability matrix to its corresponding independence distribution by means of projection operators.

COROLLARY 2.3. *Let* $Q^* = D_r^{-1/2}QD_c^{-1/2}$, *where* $\mathbf{r} > 0$, $\mathbf{c} > 0$. *Then*

$$Q^*=P_{\mathscr{L}(\mathbf{r}^{1/2})}Q^*P_{\mathscr{L}(\mathbf{c}^{1/2})}+P_{\mathscr{L}^{\perp}(\mathbf{r}^{1/2})}Q^*P_{\mathscr{L}^{\perp}(\mathbf{c}^{1/2})}. \tag{2.3}$$

Proof. Equation (2.3) follows from (2.2) if it is shown that

$$D_r^{-1/2}P^{\mathscr{L}^{\perp}(\mathbf{1})}_{\mathscr{L}(\mathbf{r})}D_r^{1/2}=P_{\mathscr{L}(\mathbf{r}^{1/2})} \tag{2.4}$$

and

$$D_c^{1/2}P^{\mathscr{L}^{\perp}(\mathbf{c})}_{\mathscr{L}(\mathbf{1})}D_c^{-1/2}=P_{\mathscr{L}(\mathbf{c}^{1/2})}. \tag{2.5}$$

The symmetry of the matrix in the l.h.s. of (2.4) follows if $D_r^{-1/2}P^{\mathscr{L}^{\perp}(\mathbf{1})}_{\mathscr{L}(\mathbf{r})}D_r^{1/2} = D_r^{1/2}P^{\mathscr{L}^{\perp}(\mathbf{r})}_{\mathscr{L}(\mathbf{1})}D_r^{-1/2}$, which holds because $P^{\mathscr{L}^{\perp}(\mathbf{1})}_{\mathscr{L}(\mathbf{r})}D_r = \mathbf{rr}'$ implies $P^{\mathscr{L}^{\perp}(\mathbf{1})}_{\mathscr{L}(\mathbf{r})}D_r = D_rP^{\mathscr{L}^{\perp}(\mathbf{r})}_{\mathscr{L}(\mathbf{1})}$. Because this matrix is also idempotent and has rank of one, to show (2.4) it is sufficient that $P^{\mathscr{L}^{\perp}(\mathbf{1})}_{\mathscr{L}(\mathbf{r})}D_r^{1/2}\mathbf{r}^{1/2} = D_r^{1/2}\mathbf{r}^{1/2}$. Analogously, (2.5) holds.

While the projection matrices in (2.2) (or in (2.3)) are unique, there are various ways to represent them by choosing different bases for $\mathscr{L}^{\perp}(\mathbf{r})$ and $\mathscr{L}^{\perp}(\mathbf{c})$. A basis (or basis matrix) for $\mathscr{L}^{\perp}(\mathbf{r})$ is an $m \times (m-1)$ full rank matrix A whose columns are orthogonal to $\mathbf{r}$, and a basis for $\mathscr{L}^{\perp}(\mathbf{c})$ is an $n \times (n-1)$ full rank matrix B whose columns are orthogonal to $\mathbf{c}$. Define

$$\Sigma_{11}=A'D_rA, \tag{2.6a}$$

$$\Sigma_{22}=B'D_cB \tag{2.6b}$$

and

$$\Sigma_{12}=A'QB(=\Sigma_{21}'). \tag{2.6c}$$

The following technical lemma permits the representation of the relevant projection matrices in terms of A, B, Σ_{11} and Σ_{22}.

LEMMA 2.4. *Suppose A and B are basis matrices for $\mathscr{L}^{\perp}(r)$ and $\mathscr{L}^{\perp}(c)$, respectively. Let $A_g = \Sigma_{11}^{-1}A'D_r$ and $B_g = \Sigma_{22}^{-1}B'D_c$. Then*

(a) *A_g is a generalized inverse of A and B_g is a generalized inverse of B;*

(b) $P^{\mathscr{L}(\mathbf{r})}_{\mathscr{L}^{\perp}(\mathbf{1})} = (AA_g)'$ *and* $P^{\mathscr{L}(\mathbf{1})}_{\mathscr{L}^{\perp}(\mathbf{c})} = (BB_g)$.

Proof. Part (a) is obvious. To show part (b), note that $(AA_g)' = (A(A'D_rA)^{-1}A'D_r)' = D_r^{1/2}P_{\mathscr{L}(D_r^{1/2}A)}D_r^{-1/2} = D_r^{1/2}P_{\mathscr{L}^{\perp}(\mathbf{r}^{1/2})}D_r^{-1/2} = I - D_r^{1/2}P_{\mathscr{L}(\mathbf{r}^{1/2})}D_r^{-1/2} = I - P^{\mathscr{L}^{\perp}(\mathbf{1})}_{\mathscr{L}(\mathbf{r})} = P^{\mathscr{L}(\mathbf{r})}_{\mathscr{L}^{\perp}(\mathbf{1})}$, with a similar result for BB_g.

THEOREM 2.5. *Let Q be an $m \times n$ probability matrix with* $\mathbf{r} > 0$ *and* $\mathbf{c} > 0$ *and let A and B be basis matrices for* $\mathscr{L}^{\perp}(\mathbf{r})$ *and* $\mathscr{L}^{\perp}(\mathbf{c})$, *respectively. Then*

$$Q=\mathbf{rc}'+D_rA\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}B'D_c. \tag{2.7}$$

Proof. The proof follows directly from Lemma 2.4 and Remark 2.2.

As noted in § 1, the eigenvalues of $Q^{*\prime}Q^*$ or $(Q^*Q^{*\prime})$ play an important role. In the following theorem, we show that the eigenvalues of $Q^{*\prime}Q^*$ (or $Q^*Q^{*\prime}$) can be obtained from $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (or $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$), no matter the choice of basis matrices A and B.

THEOREM 2.6. (i) *The eigenvalues of* $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (*or* $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$) *do not depend on the explicit choice of the basis matrices A and B.*

(ii) *The eigenvalues of* $Q^*Q^{*\prime}$ (*or* $Q^{*\prime}Q^*$) *are* 1, *and the eigenvalues of* $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (*or* $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$).

Proof. To show (i), let A_1 and A_2 be arbitrary basis matrices for $\mathscr{L}^{\perp}(\mathbf{r})$, and B_1 and B_2 for $\mathscr{L}^{\perp}(\mathbf{c})$. Then there exists $(m-1)\times(m-1)$ and $(n-1)\times(n-1)$ nonsingular matrices G and H, respectively, such that $A_1 = A_2G$ and $B_1 = B_2H$. Hence

$$[\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]_{A_2,B_2} = G^{-1}[\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]_{A_1,B_1}G, \tag{2.8}$$

where $[\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]_{A,B}$ denotes the matrix $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ derived from the bases A, B. The result now follows from (2.8).

To show (ii), let $A^* = [D_r^{1/2}\mathbf{1}_m : D_r^{1/2}A]$ and $B^* = [D_c^{1/2}\mathbf{1}_n : D_c^{1/2}B]$. Then

$$A^{*\prime}A^* = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{11} \end{bmatrix}, \quad B^{*\prime}B^* = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{22} \end{bmatrix} \quad \text{and} \quad A^{*\prime}Q^*B^* = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{12} \end{bmatrix}.$$

Furthermore,

$$\begin{aligned} Q^*Q^{*\prime} &= A^*\begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix} B^{*\prime}Q^{*\prime}(A^*A^{*-1}) \\ &= A^*\begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{bmatrix} A^{*-1}, \end{aligned}$$

so that the result now follows.

Note that because the eigenvalues of $Q^*Q^{*\prime}$ are less than or equal to 1 (Lancaster (1969, Corollary 1, p. 90)), the eigenvalues of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ have the same property.

THEOREM 2.7. *For every choice of basis matrices for A and B, the matrix*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

is nonnegative definite, where Σ_{11}, Σ_{12}, Σ_{22} *and* Σ_{21} *are defined by* (2.6).

Proof. The nonnegative definiteness of Σ is equivalent to the nonnegative definiteness of

$$\begin{bmatrix} I & Q^* \\ Q^{*\prime} & I \end{bmatrix}.$$

Because the eigenvalues of $Q^{*\prime}Q^*$ are between 0 and 1 (e.g., Lancaster (1969, Corollary 1, p. 90)), the result follows.

In § 3 we identify explicitly a set of $m + n - 2$ random variables $(W_1, \cdots, W_{m-1}, Z_1, \cdots, Z_{n-1})' = (\mathbf{W}' : \mathbf{Z}')'$ whose covariance matrix is Σ. In light of Theorems 2.6 and 2.7, we note that the canonical correlations between X, Y in the sense of Fisher and Lancaster can be viewed as the canonical correlations between $\mathbf{W}$ and $\mathbf{Z}$ in the sense of Hotelling (see Anderson (1984, Chap. 12)).

3. Derivation of the canonical and quadrant dependence decompositions. In this section, we demonstrate that the version of the projection theorem given by Theorem 2.5, in fact, yields for a suitable choice of the basis matrices A and B Fisher's canonical decomposition, and for another choice the quadrant dependence decomposition. The former result is stated in Theorem 3.1 and the latter in Theorem 3.3.

THEOREM 3.1. *Let* Q *be an* $m \times n$ *probability matrix of rank* k *with* $\mathbf{r} > 0$ *and* $\mathbf{c} > 0$. *Let* $\Gamma = (\boldsymbol{\Gamma}_1 : \Gamma_2)$ *and* $G = (\mathbf{G}_1 : G_2)$ *denote, respectively,* $m \times m$ *and* $n \times n$ *orthogonal matrices such that* $Q^*Q^{*\prime} = \Gamma D_1\Gamma'$ *and* $Q^{*\prime}Q^* = GD_2G'$, *where* $D_1 =$ Diag $(1, \rho_1^2, \cdots, \rho_{k-1}^2, 0, \cdots, 0)$, $D_2 =$ Diag $(1, \rho_1^2, \cdots, \rho_{k-1}^2, 0, \cdots, 0)$, *and* $1 \geqq$

$\rho_1^2 \geqq \cdots \geqq \rho_{k-1}^2$ are the nonzero eigenvalues of $Q^*Q^{*\prime}$ (or $Q^{*\prime}Q^*$). Then $A_0 = D_r^{-1/2}\Gamma_2$ and $B_0 = D_c^{-1/2}G_2$ are basis matrices for $\mathscr{L}^\perp(\mathbf{r})$ and $\mathscr{L}^\perp(\mathbf{c})$, respectively. Moreover the decomposition

$$Q = \mathbf{rc}' + D_r A_0 \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}B_0' D_c \tag{3.1}$$

yields the canonical decomposition, where Σ_{11}, Σ_{12} and Σ_{22} are defined by (2.6).

Proof. The vector $D_r^{1/2}\mathbf{1}_m$ is clearly an eigenvector of $Q^*Q^{*\prime}$ corresponding to the eigenvalue of 1. Therefore, $A_0'\mathbf{r} = \Gamma_2' D_r^{-1/2} D_r \mathbf{1}_m = \Gamma_2'\boldsymbol{\Gamma}_1 = \mathbf{0}$ which implies from the orthogonality of Γ that A_0 is a basis matrix for $\mathscr{L}^\perp(\mathbf{r})$. Similarly, B_0 is a basis matrix for $\mathscr{L}^\perp(\mathbf{c})$. Moreover, $\Sigma_{11} = A_0' D_r A_0 = \Gamma_2'\Gamma_2 = I_{m-1}$; $\Sigma_{22} = I_{n-1}$; and

$$\Sigma_{12} = A_0' Q B_0 = \Gamma_2' Q^* G_2 = \begin{bmatrix} \text{Diag}\,(\rho_1, \cdots, \rho_k) & 0 \\ 0 & 0 \end{bmatrix},$$

so that (3.1) yields (1.2).

In order to derive the quadrant dependence decomposition from the projection decomposition, we require the matrices $\bar{A}$ and $\bar{B}$ to be defined by

$$\bar{A} = (I - J_m D_r)\psi_m \tag{3.2a}$$

and

$$\bar{B} = (I - J_n D_c)\psi_n \tag{3.2b}$$

where the $p \times (p-1)$ matrix ψ_p is defined by $(\Sigma_{k>1}\,\mathbf{e}_k : \cdots : \Sigma_{k>p-1}\,\mathbf{e}_k)$.

Three straightforwardly provable properties of $\bar{A}$ and $\bar{B}$ are given below.

LEMMA 3.2. *Suppose $\bar{A}$ and $\bar{B}$ are defined in* (3.2). *Then*

(i) $\bar{A}'Q\bar{B} = \psi_m'(Q - \mathbf{rc}')\psi_n$.

(ii) *A generalized inverse of $\bar{A}$ is Δ_m' and of $\bar{B}$ is Δ_n', where Δ_p is defined by* (1.4).

(iii) $\Delta_m' = \Sigma_{11}^{-1}\bar{A}'D_r$ *and* $\Delta_n' = \Sigma_{22}^{-1}\bar{B}'D_c$, *where* $\Sigma_{11} = \bar{A}'D_r\bar{A}$ *and* $\Sigma_{22} = \bar{B}'D_c\bar{B}$.

THEOREM 3.3. *Let Q be an $m \times n$ probability matrix with $\mathbf{r} > 0$ and $\mathbf{c} > 0$. Let $\bar{A}$ and $\bar{B}$ be defined by* (3.2). *Then $\bar{A}$ and $\bar{B}$ are basis matrices for $\mathscr{L}^\perp(\mathbf{r})$ and $\mathscr{L}^\perp(\mathbf{c})$, respectively. Moreover, the decomposition*

$$Q = \mathbf{rc}' + D_r\bar{A}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\bar{B}'D_c \tag{3.3}$$

yields the quadrant dependence decomposition, where Σ_{11}, Σ_{12} and Σ_{22} are defined by (2.6).

Proof. Note that $\Delta_m'\bar{A} = \Delta_m'\psi_m = I_{m-1}$ which implies that the rank of $\bar{A} = m - 1$. Also $\bar{A}'\mathbf{r} = \psi_m'(I - D_r J_m)\mathbf{r} = \psi_m'(\mathbf{r} - \mathbf{r}) = \mathbf{0}$. Hence $\bar{A}$ is a basis matrix for $\mathscr{L}^\perp(\mathbf{r})$, and similarly $\bar{B}$ is a basis matrix for $\mathscr{L}^\perp(\mathbf{c})$. Theorem 2.5 along with Lemma 3.2 yields the following result:

$$Q = \mathbf{rc}' + \Delta_m\Sigma_{12}\Delta_n' \tag{3.4}$$

where $\Sigma_{12} = \psi_m'(Q - \mathbf{rc}')\psi_n$. But $\psi_m' Q\psi_n = \bar{H}$, $\psi_m'\mathbf{r} = \bar{\mathbf{F}}$, and $\mathbf{c}'\psi_n = \bar{\mathbf{G}}'$, so that the quadrant dependence decomposition of (1.3) now follows.

Simple algebra yields that for the quadrant dependence decomposition $\Sigma_{11} = \{f_{ij}\}$, where $f_{ij} = (1 - \bar{F}_{\min(i,j)})\bar{F}_{\max(i,j)}$ and $\Sigma_{22} = \{g_{ij}\}$, where $g_{ij} = (1 - \bar{G}_{\min(i,j)})\bar{G}_{\max(i,j)}$. Viewing Σ_{11} and Σ_{22} as functions of $\mathbf{r}$ and $\mathbf{c}$, respectively, we denote this by $\bar{\Sigma}_{11}(\mathbf{r})$ and $\bar{\Sigma}_{22}(\mathbf{c})$.

COROLLARY 3.4. *The squares of the canonical correlations of Q are the eigenvalues of $\bar{\Sigma}_{11}^{-1}(\mathbf{r})(\bar{H} - \bar{\mathbf{F}}\bar{\mathbf{G}}')\bar{\Sigma}_{22}^{-1}(\mathbf{c})(\bar{H}' - \bar{\mathbf{G}}\bar{\mathbf{F}}')$.*

Proof. This follows immediately from Theorem 3.3 and Theorem 2.6.

As noted in Theorem 2.7, the matrix

$$\bar{\Sigma}(Q) \equiv \begin{bmatrix} \bar{\Sigma}_{11}(\mathbf{r}) & \bar{\Sigma}_{12}(Q) \\ \bar{\Sigma}_{21}(Q) & \bar{\Sigma}_{22}(\mathbf{c}) \end{bmatrix} \tag{3.5}$$

where $\bar{\Sigma}_{12}(Q) = \bar{H} - \bar{\mathbf{F}}\bar{\mathbf{G}}'$, is nonnegative definite for each probability matrix Q. An interesting and useful converse is to find conditions on an arbitrary $(m + n - 2) \times (m + n - 2)$ nonnegative definite matrix Σ_0 so that for some $m \times n$ probability matrix Q_0, we have $\Sigma_0 = \bar{\Sigma}(Q_0)$. Partition

$$\Sigma_0 \text{ as } \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

so that Σ_{11} is $(m - 1) \times (m - 1)$. Clearly, there must exist $\mathbf{r} > 0$, with $\Sigma r_i = 1$ and $\mathbf{c} > 0$ with $\Sigma c_j = 1$, so that $\Sigma_{11} = \bar{\Sigma}_{11}(\mathbf{r})$ and $\Sigma_{22} = \bar{\Sigma}_{22}(\mathbf{c})$. Then for such $\mathbf{r}$ and $\mathbf{c}$, we must have from (3.4) that $\mathbf{rc}' + \Delta_m\Sigma_{12}\Delta_n'$ has all nonnegative elements. This is summarized in the following lemma.

LEMMA 3.5. *Let*

$$\Sigma_0 = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

be a nonnegative definite matrix. If there exists $\mathbf{r} > 0$, *with* $\Sigma r_i = 1$, *and* $\mathbf{c} > 0$, *with* $\Sigma c_j = 1$, *such that* $\Sigma_{11} = \bar{\Sigma}_{11}(\mathbf{r})$ *and* $\Sigma_{22} = \bar{\Sigma}_{22}(\mathbf{c})$, *and if* $\mathbf{rc}' + \Delta_m\Sigma_{12}\Delta_n' \geqq 0$, *then there exists a probability matrix* Q_0 *such that* $\Sigma_0 = \bar{\Sigma}(Q_0)$, *where* $\bar{\Sigma}(Q)$ *is given by* (3.5).

In some ways, Lemma 3.5 can be viewed as an analogue for the quadrant dependence decomposition of results like those of Tyan and Thomas (1975) obtained for the canonical decomposition in the more general continuous case.

We conclude this section by identifying $m + n - 2$ random variables whose covariance matrix is $\bar{\Sigma}$. For a given p.m.f. matrix Q, define the nondecreasing functions $W_1, \cdots, W_{m-1}$ and $Z_1, \cdots, Z_{n-1}$ by

$$W_k(x) = -\bar{F}_k I_{(-\infty, x_k]}(x) + F_k I_{(x_k, \infty)}(x), \qquad k = 1, \cdots, m-1 \tag{3.6a}$$

and

$$Z_l(y) = -\bar{G}_l I_{(-\infty, y_l]}(y) + G_l I_{(y_l, \infty)}(y), \qquad l = 1, \cdots, n-1 \tag{3.6b}$$

where $I_{(a,b]}(u)$ is the indicator function of $(a, b]$. The following theorem is straightforward to prove.

THEOREM 3.6. *Let Q be an $m \times n$ probability matrix with* $\mathbf{r} > 0$ *and* $\mathbf{c} > 0$. *Define the* $(m + n - 2)$-*dimensional random vector* $\mathbf{U}$ *by*

$$\mathbf{U} = (W_1(X), \cdots, W_{m-1}(X), Z_1(Y), \cdots, Z_{n-1}(Y))'$$

where $\{W_k(X)\}$ *and* $\{Z_l(Y)\}$ *are given by* (3.6). *Then* $E(\mathbf{U}) = \mathbf{0}$ *and* $E(\mathbf{UU}') = \bar{\Sigma}$.

4. Applications. We now consider some further applications of the previous results, in particular, of the basis matrices $\bar{A}$ and $\bar{B}$ of (3.2). A vector $\mathbf{x} = (x_1, \cdots, x_p)'$ is said to be increasing if $x_1 \leqq \cdots \leqq x_p$, i.e., if $\Delta_p'\mathbf{x} \geqq 0$, and decreasing if $-\mathbf{x}$ is increasing. The basis matrices $\bar{A}$ and $\bar{B}$ provide a convenient way of representing a vector's being increasing, as is noted in the following lemma whose proof is straightforward.

LEMMA 4.1. *Every increasing* (*decreasing*) *vector* $\mathbf{x} \in \mathscr{L}^{\perp}(\mathbf{r})$ *can be written as* $\mathbf{x} = \bar{A}\alpha$ *if and only if* $\alpha \geqq (\leqq) 0$.

Kimeldorf, May and Sampson (1982) introduced the concordant monotone correlation coefficient (CMC), which for two ordinal variables X and Y with probability matrix Q, is defined by

$$\text{(4.1)} \qquad \text{CMC}\,(Q) = \max \mathbf{x}'Q\mathbf{y}$$

where the maximum is taken over all increasing vectors $\mathbf{x} \in \mathscr{L}^{\perp}(\mathbf{r})$ and $\mathbf{y} \in \mathscr{L}^{\perp}(\mathbf{c})$, (i.e., $EX = EY = 0$) such that $\mathbf{x}'D_r\mathbf{x} = 1$ and $\mathbf{y}'D_c\mathbf{y} = 1$ (i.e., $EX^2 = EY^2 = 1$). Using Lemma 4.1 one can readily show that

$$\text{(4.2)} \qquad \text{CMC}\,(Q) = \max \alpha'\bar{\Sigma}_{12}(Q)\beta$$

where the maximum is taken over all $\alpha \geqq 0$, $\beta \geqq 0$ with $\alpha'\bar{\Sigma}_{11}(\mathbf{r})\alpha = 1$ and $\beta'\bar{\Sigma}_{22}(\mathbf{c})\beta = 1$. Formulation of the concordant monotone correlation optimization problem as (4.2) instead of (4.1) has the apparent benefit of reduced dimensionality and nonnegativity constraints in place of the monotonicity constraints.

If we remove the constraint of $\alpha \geqq 0$ and $\beta \geqq 0$ in the optimization problem of (4.2), the resulting measure is $\rho'(Q)$, the maximal correlation coefficient. In fact, to compute the maximal correlation coefficient, it suffices to replace in (4.2), $\bar{\Sigma}(Q)$ by Σ_{11}, Σ_{12}, Σ_{22}, where these three matrices are computed from arbitrary basis matrices using (2.6). The following theorem and its corollary provide an approach to finding $\rho'(Q)$ and conditions for having CMC $(Q) = \rho'(Q)$.

THEOREM 4.2. *Let Q be an $m \times n$ probability matrix with* $\mathbf{r} > 0$, $\mathbf{c} > 0$ *and let A and B be basis matrices for* $\mathscr{L}^{\perp}(\mathbf{r})$ *and* $\mathscr{L}^{\perp}(\mathbf{c})$, *respectively. Then* $\rho'(Q) = \alpha'_*\Sigma_{12}\beta_*$, *where α_* and β_* are any vectors satisfying*

$$\text{(4.3)} \qquad \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha_* = (\rho'(Q))^2\alpha_*,$$

$$\text{(4.4)} \qquad \alpha'_*\Sigma_{11}\alpha_* = 1$$

and

$$\text{(4.5)} \qquad \beta_* = (\rho'(Q))^{-1}\Sigma_{22}^{-1}\Sigma_{21}\alpha_*$$

where Σ_{11}, Σ_{12}, Σ_{22} *are given by* (2.6) *and $\rho'(Q)$ is assumed positive.*

Proof. Suppose α_* and β_* satisfy (4.3), (4.4) and (4.5). Then

$$\begin{aligned} \alpha'_*\Sigma_{12}\beta_* &= (\rho'(Q))^{-1}\alpha'_*\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha_* \\ &= \rho'(Q)\alpha'_*\Sigma_{11}\alpha_* \\ &= \rho'(Q). \end{aligned}$$

COROLLARY 4.3. *If* $\bar{\Sigma}_{11}^{-1}(\mathbf{r})\bar{\Sigma}_{12}(Q) \geqq 0$ *and* $\bar{\Sigma}_{22}^{-1}(\mathbf{c})\bar{\Sigma}_{21}(Q) \geqq 0$, *then* CMC $(Q) = \rho'(Q)$.

Proof. It follows that $\bar{C} = \bar{\Sigma}_{11}^{-1}(\mathbf{r})\bar{\Sigma}_{12}(Q)\bar{\Sigma}_{22}^{-1}(\mathbf{c})\bar{\Sigma}_{21}(Q) \geqq 0$. Let $\alpha^{(1)}$ be the eigenvector of $\bar{C}$ corresponding to the largest root $\rho_1^2 = (\rho'(Q))^2$ such that $\alpha^{(1)\prime}\Sigma_{11}(\mathbf{r})\alpha^{(1)} = 1$. It follows, e.g., Gantmacher (1959, p. 66), that $\alpha^{(1)} \geqq 0$. Let $\beta^{(1)} = (\rho'(Q))^{-1}\bar{\Sigma}_{22}^{-1}(\mathbf{c})\bar{\Sigma}_{21}(Q)\alpha^{(1)}$, so that $\beta^{(1)} \geqq 0$ and $\beta^{(1)\prime}\Sigma_{22}(\mathbf{c})\beta^{(1)} = 1$. Hence, from Theorem 4.2, and (4.2), it follows that $\rho'(Q) = \alpha'_*\bar{\Sigma}_{12}(Q)\beta_* = \text{CMC}\,(Q)$.

Corollary 4.3 is implicitly contained in the results of Schriever (1983).

To see that the conditions of Corollary 4.3 are not necessary, let

$$Q = \begin{bmatrix} 2/8 & 2/8 & 0 \\ 2/8 & 1/8 & 0 \\ 0 & 0 & 1/8 \end{bmatrix}.$$

Then $\rho'(Q) = \mathrm{CMC}\,(Q) = 1$, but

$$\bar{\Sigma}_{11}^{-1}(\mathbf{r})\bar{\Sigma}_{12}(Q) = \bar{\Sigma}_{22}^{-1}(\mathbf{c})\bar{\Sigma}_{21}(Q) = \begin{bmatrix} -1/6 & 0 \\ 4/6 & 1 \end{bmatrix}.$$

We conclude by establishing a connection between a problem related to the tradition canonical correlation analysis of the multivariate normal and the CMC. Suppose

$$(\mathbf{X}_1' : \mathbf{X}_2')' \sim N\left(\mathbf{0}, \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}\right)$$

where $\mathbf{X}_1$ is $(m-1) \times 1$ and $\mathbf{X}_2$ is $(n-1) \times 1$. Define $\rho^+(\mathbf{X}_1, \mathbf{X}_2) = \sup \rho(\alpha'\mathbf{X}_1, \beta'\mathbf{X}_2)$, subject to $\mathrm{Var}\,(\alpha'\mathbf{X}_1) > 0$, $\mathrm{Var}\,(\beta'\mathbf{X}_2) > 0$ and $\alpha \geqq 0$, $\beta \geqq 0$. Following the terminology of Waterman (1974), we call ρ^+ the *nonnegative canonical correlation coefficient.* This optimization problem does not appear to be addressed in the statistical literature, and seems to be fairly difficult to solve for general $\tau_{11}, \tau_{12}, \tau_{22}$. However, if there exists a pmf matrix Q_+ so that

$$\begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix} = \bar{\Sigma}(Q_+)$$

(see Lemma 3.5) then clearly $\rho^+ = \mathrm{CMC}\,(Q_+)$. In this case the program MONCOR described by Kimeldorf, May and Sampson (1982) would permit the computation of ρ^+. Also using the argument of the proof of Corollary 4.3, and standard results (e.g., Anderson (1984, Chap. 12)), we can establish the following theorem.

THEOREM 4.4. *Let*

$$(\mathbf{X}_1' : \mathbf{X}_2')' \sim N\left(\mathbf{0}, \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}\right).$$

If $\tau_{11}^{-1}\tau_{12} \geqq 0$ *and* $\tau_{22}^{-1}\tau_{21} \geqq 0$, *then* $\rho^+(\mathbf{X}_1, \mathbf{X}_2)$ *is Hotelling's canonical correlation coefficient between* $\mathbf{X}_1$ *and* $\mathbf{X}_2$.

REFERENCES

T. W. ANDERSON (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley, New York.

J. P. BENZECRI (1973), *L'analyse des donnees* II: *l'analyse des correspondences*, Dunod, Paris.

R. A. FISHER (1940), *The precision of discriminant functions*, Ann. Eugen. London, 10, pp. 422–429.

F. R. GANTMACHER (1959), *The Theory of Matrices*, Chelsea, New York.

Z. GILULA (1984), *On some similarities between canonical correlation models and latent class models for two-way contingency tables*, Biometrika, 71, pp. 523–529.

H. O. HIRSCHFELD (1935), *A connection between correlation and contingency*, Proc. Camb. Philos. Soc., 31, pp. 520–524.

G. KIMELDORF AND A. R. SAMPSON (1978), *Monotone dependence*, Ann. Statist., 6, pp. 895–903.

G. KIMELDORF, J. MAY AND A. R. SAMPSON (1982), *Concordant and discordant monotone correlations and their evaluation by nonlinear optimization*, in Optimization in Statistics, S. H. Zanakis and J. S. Rustagi, eds., TIMS Studies Management Sci., 19, pp. 117–130.

H. O. LANCASTER (1969), *The Chi-squared Distribution*, John Wiley, New York.

——— (1958), *The structure of bivariate distributions*, Ann. Math. Statist., 29, pp. 719–736.

E. L. LEHMANN (1966), *Some concepts of dependence*, Ann. Math. Statist., 37, pp. 1137–1153.

K. MAUNG (1941), *Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children*, Ann. Eugen. London, 11, pp. 189–223.

S. NISHISATO AND P. S. ARRI (1975), *Nonlinear programming approach to optimal scaling of partially ordered categories*, Psychometrika, 40, pp. 525–548.

S. NISHISATO (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*, University of Toronto Press, Toronto, Ontario, Canada.

O. V. SARMANOV (1958a), *The maximal correlation coefficient* (*symmetric case*), Dokl. Akad. Nauk. SSSR, 120, pp. 715–718. (In Russian.) Selected Transl. Math. Statist. Probab., 4, pp. 271–275. (In English.)

——— (1958b), *The maximal correlation coefficient* (*non-symmetric case*), Dokl. Akad. Nauk. SSSR, 121, pp. 52–55. (In Russian.) Selected Transl. Math. Statist. Probab., 4, pp. 207–210. (In English.)

B. F. SCHRIEVER (1985), *Order dependence*, Ph.D. Dissertation, Free University of Amsterdam,

——— (1983), *Scaling of order dependent categorical variables with correspondence analysis*, Internat. Statis. Rev., 51, pp. 225–238.

Y. L. TONG (1980), *Probability Inequalities in Multivariate Distributions*, Academic Press, New York.

S. TYAN AND J. B. THOMAS (1975), *Characterization of a class of bivariate distributions*, J. Multivariate Anal., 5, pp. 227–235.

M. S. WATERMAN (1974), *A restricted least squares problem*, Technometrics, 16, pp. 135–136.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 3, July 1987

017

OPTIMAL ASSIGNMENTS FOR CONSECUTIVE-2 GRAPHS*

D. Z. DU† AND F. K. HWANG‡

Abstract. Let G represent the graph structure of a system of components each of which can either work or fail. Suppose that the system itself fails if and only if two adjacent vertices both fail; then G is called a consecutive-2 graph. Given a set of probabilities $p = \{p_1, \cdots, p_n\}$ where n is the number of vertices in G, the problem is to assign p_i to the n vertices to minimize the probability of the system failing. Previous literature has dealt with the case that G consists of lines and cycles. Here we give some results applicable to general graphs. We also discuss the conditions for G to have an optimal assignment which depends only on the ranks of p_i.

Key words. consecutive-2-out-of-n system, consecutive-2 graph, reliability

AMS(MOS) subject classification. 90B25

1. Introduction. Let $G = (V, E)$ be a graph with vertex set $V = \{v_1, \cdots, v_n\}$ and an edge set E. Suppose that each vertex can either work or fail. G is called a *consecutive*-2 *graph* provided that G fails if and only if two adjacent vertices both fail. Let $p = \{p_1 \leqq p_2 \leqq \cdots \leqq p_n\}$ denote a set of n probabilities. An *assignment* P is a one-to-one mapping from p to V, i.e., p_i, when assigned to v_j, will be interpreted as the probability that v_j works (does not fail). The problem is to find the assignment which minimizes the probability of G failing.

An example of a potential application of this model is the storage of m pieces of data into n memory units which can occasionally lose their contents. To prevent loss of data we decide to duplicate each piece of data and store it in two distinct memory units. So G is the graph with memory units as vertices and an edge between two memory units if there exists a piece of data stored in both units. The storage program is considered failed if any piece of data is lost. The problem is to assign the n memory units to a fixed set of n environments (an environment can affect the performance of a memory unit) to minimize the failure probability of the storage program.

Optimal assignments have been obtained for consecutive-2 graphs consisting of lines and cycles [1]–[4]. In this paper we will give some results applicable to general graphs which help to narrow down the candidates for optimal assignments. We also discuss the conditions for G to have an optimal assignment which depends only on the ranks of p_i.

2. The main results. Consider an automorphism θ of G, that is, θ: $V \rightarrow V$ is one-to-one, onto and such that $[\theta(u), \theta(v)] \in E$ if and only if $[u, v] \in E$. If $\theta^2 = 1$, then $\theta(u) = v$ is an equivalence relation on V and each equivalence class contains one or two vertices. Let $\tilde{u}$ denote the equivalence class containing u. Define a new graph $G_\theta = (V_\theta, E_\theta)$ where $V_\theta = \{\tilde{u} \mid u \in V\}$ and an edge $[\tilde{u}, \tilde{v}]$ exists for every pair of edges $[u, v] \in E$ and $[\theta(u), \theta(v)] \in E$. Note that if $\tilde{u} = \{u\}$ and $\tilde{v} = \{v, v'\}$ with $[u, v] \in E$ and $[u, v'] \in E$, then there is only one edge between $\tilde{u}$ and $\tilde{v}$ since the two pairs $(\{u, v\}, \{\theta(u), \theta(v)\})$ and $(\{u, v'\}, \{\theta(u), \theta(v')\})$ are identical.

* Received by the editors January 17, 1985; accepted for publication (in revised form) November 26, 1986.

† University of California, Santa Barbara, California 93106. This author is on leave from the Institute of Applied Mathematics, Academia Sinica, Beijing, China.

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

θ is called *regular* if it satisfies the following three conditions:

(i) $\theta^2 = 1$.

(ii) G_θ has no multiple edges, i.e., there do not exist four distinct edges $[u, v]$, $[u, \theta(v)]$, $[\theta(u), v]$, $[\theta(u), \theta(v)]$ in E.

(iii) Let $(\tilde{u}_0, \tilde{u}_1, \cdots, \tilde{u}_m, \tilde{u}_{m+1} = \tilde{u}_0)$ denote any cycle in G_θ and let u_i be an arbitrary element of $\tilde{u}_i$, $i = 0, 1, \cdots, m$. Then $|(U^m_{i=0}[u_i, u_{i+1}]) \cap E|$ is always an even number.

An edge $[\tilde{u}, \tilde{v}] \in E_\theta$ is called *normal* (or *singular*) with respect to P_θ if $[u, v] \in E$ and $[P(u) - P(\theta(u))][P(v) - P(\theta(v))] \leqq 0$ (or >0).

THEOREM 1. *Let θ be a regular automorphism of G. Then for any assignment P there exists an assignment P_θ such that*

(i) $\{P_\theta(v), P_\theta(\theta(v))\} = \{P(v), P(\theta(v))\}$ *for all* $v \in V$;

(ii) *G_θ has no singular edge with respect to P_θ.*

Proof. If G_θ has no singular edge with respect to P, set $P_\theta = P$ and we are through. Otherwise, let $[\tilde{u}_0, \tilde{v}_0]$ denote a singular edge of G_θ with respect to P. Set $E'_\theta = \{[\tilde{u}, \tilde{v}] \in E_\theta | [\tilde{u}, \tilde{v}]$ *is normal with respect to* $P\}$ and $G'_\theta = (V_\theta, E'_\theta)$. Then $\tilde{u}_0$ and $\tilde{v}_0$ are not connected in G'_θ. Suppose to the contrary that there exists a path

$$(\tilde{u}_0, \tilde{u}_1, \cdots, \tilde{u}_m = \tilde{v}_0).$$

Let $u_i \in \tilde{u}_i$ for $i = 0, 1, \cdots, m$ such that $[u_i, u_{i+1}] \in E$ for $i = 0, 1, \cdots, m - 1$. Without loss of generality, assume $P(u_0) \geqq P(\theta(u_0))$. Since $[\tilde{u}_i, \tilde{u}_{i-1}]$ are all normal, we have $P(u_m) \geqq P(\theta(u_m))$ for m even and $P(u_m) \leqq P(\theta(u_m))$ for m odd. Since $[\tilde{u}_0, \tilde{u}_m]$ is singular, we have $[u_0, u_m] \in E$ for m even and $[u_0, u_m] \notin E$ for m odd. In either case, $|(U^m_{i=0}[u_i, u_{i+1}]) \cap E|$ is odd, contradicting condition (iii) of a regular automorphism.

Define C to be the component of G'_θ containing $\tilde{u}_0$. Define:

$$P_1(u) = \begin{cases} P(u) & \text{if } u \in C, \\ P(\theta(u)) & \text{if } u \notin C. \end{cases}$$

Then every edge of G_θ normal with respect to P is also normal with respect to P_1 since the components of G'_θ remain intact. But P_1 contains one additional normal edge $[\tilde{u}_0, \tilde{u}_m]$. Replace P by P_1 in the above argument and proceed accordingly, eventually we obtain P_θ with respect to which G_θ has no singular edge. □

Let $P\{G, A\}$ denote the joint probability of the graph G working and the event A occurring under the assignment P. Let $P(A)$ denote the probability of the event A occurring under the assignment P. For U a set of vertices in G, $G - U$ denotes the graph obtained from G by deleting U and all edges incident to it. Consider $\tilde{u} = \{u, u'\} \in V_\theta$. Suppose that $u \neq u'$ and $P(u) \geqq P(u')$. Define $\tilde{u}^l(\tilde{u}^s)$ to be the event that $u(u')$ works and $u'(u)$ fails. For $u = u'$ define $\tilde{u}^l = \tilde{u}^s =$ an impossible event, i.e., $P(\tilde{u}^l) = P(\tilde{u}^s) = 0$. We also define $\bar{l} = s$ and $\bar{s} = l$.

THEOREM 2. *Let θ be a regular automorphism of G and P_θ an assignment such that G_θ has no singular edges. Then for $U \subseteq V_\theta$ and $x(\tilde{u}) = l$ or s we have*:

$$\text{(i)}\quad P_\theta\Bigl(G, \bigcap_{\tilde{u} \in U} \tilde{u}^l\Bigr) \geqq P\Bigl(G, \bigcap_{\tilde{u} \in U} \tilde{u}^{x(\tilde{u})}\Bigr).$$

$$\text{(ii)}\quad P_\theta\Bigl(G, \bigcap_{\tilde{u} \in U} \tilde{u}^l\Bigr) + P_\theta\Bigl(G, \bigcap_{\tilde{u} \in U} \tilde{u}^s\Bigr) \geqq P\Bigl(G, \bigcap_{\tilde{u} \in U} \tilde{u}^{x(\tilde{u})}\Bigr) + P\Bigl(G, \bigcap_{\tilde{u} \in U} \tilde{u}^{\bar{x}(\tilde{u})}\Bigr).$$

Proof. We prove Theorem 2 by induction on $|V_\theta|$. The proof is trivial for $|V_\theta| = 1$. Consider general $|V_\theta| = m > 1$. For any $\tilde{v}_0 \in V_\theta$ define $\alpha(\tilde{v}_0) = \{\tilde{v} \mid \tilde{v} \notin U, [\tilde{v}, \tilde{v}_0] \in E_\theta\}$ and $d(v_0) = \{v \in \tilde{v} \mid \tilde{v} \in \alpha(\tilde{v}_0)\}$. For any $\tilde{W} \subseteq V_\theta$ define $W = \{v \in V \mid \tilde{v} \in \tilde{W}\}$.

Case 1. $U \neq \varnothing$. Suppose that $\tilde{v}_0 = \{v_0, v'_0\} \in U$. To prove (i), we have

$$
\begin{aligned}
P_\theta\Bigg(G, \bigcap_{\tilde{u}\in U} \tilde{u}^l\Bigg) &= P_\theta(\tilde{v}_0^l) \sum_{\tilde{W}\subseteq \alpha(\tilde{v}_0)} \Bigg(\prod_{w\in W} P_\theta(w)\Bigg) P_\theta\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^l\Bigg)\\
&\geqq P(\tilde{v}_0^{x(\tilde{v}_0)}) \sum_{\tilde{W}\subseteq \alpha(\tilde{v}_0)} \Bigg(\prod_{w\in W} P(w)\Bigg)\\
&\qquad \cdot P\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^{y(\tilde{u})}\Bigg)\\
&= P\Bigg(G, \bigcap_{\tilde{u}\in U} \tilde{u}^{x(\tilde{u})}\Bigg)
\end{aligned}
\tag{1}
$$

where

$$
\begin{aligned}
y(\tilde{u}) &= x(\tilde{v}_0) \quad \text{if either } \tilde{u}\in\alpha(v_0) \text{ and } [\tilde{v}_0, \tilde{u}] \text{ is normal with respect to } P, \textit{ or } \tilde{u}\in U\\
&= \bar{x}(\tilde{v}_0) \quad \text{if } \tilde{u}\in\alpha(v_0) \text{ and } [\tilde{v}_0, \tilde{u}] \text{ is singular with respect to } P.
\end{aligned}
$$

The inequality is obtained by using the induction hypothesis (i) and noting that

$$P(\tilde{v}_0^l) \geqq P(\tilde{v}_0^{x(\tilde{v}_0)}).$$

To prove (ii) we assume without loss of generality that $x(\tilde{v}_0) = l$ from the symmetry of (ii). Then

$$
\begin{aligned}
&P_\theta\Bigg(G, \bigcap_{u\in U} \tilde{u}^l\Bigg) + P_\theta\Bigg(G, \bigcap_{u\in U} \tilde{u}^s\Bigg)\\
&= P_\theta(\tilde{v}_0^l) \sum_{\tilde{W}\subseteq \alpha(\tilde{v}_0)} \Bigg(\prod_{w\in W} P_\theta(w)\Bigg) P_\theta\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^l\Bigg)\\
&\quad + P_\theta(\tilde{v}_0^s) \sum_{\tilde{W}\subseteq \alpha(\tilde{v}_0)} \Bigg(\prod_{w\in W} P_\theta(w)\Bigg) P_\theta\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^s\Bigg)\\
&= [P_\theta(\tilde{v}_0^l) - P_\theta(\tilde{v}_0^s)] \sum_{\tilde{W}\subseteq \alpha(\tilde{v}_0)} \Bigg(\prod_{w\in W} P_\theta(w)\Bigg) P_\theta\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^l\Bigg)\\
&\quad + P_\theta(\tilde{v}_0^s) \sum_{\tilde{W}\subseteq \alpha(v_0,v'_0)} \Bigg(\prod_{w\in W} P_\theta(w)\Bigg) \Bigg[P_\theta\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^l\Bigg)\\
&\qquad + P_\theta\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^s\Bigg)\Bigg]\\
&\geqq [P(\tilde{v}_0^l) - P(\tilde{v}_0^s)] \sum_{\tilde{W}\subseteq \alpha(v_0,v'_0)} \Bigg(\prod_{w\in W} P(w)\Bigg) P\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^{y(\tilde{u})}\Bigg)\\
&\quad + P(\tilde{v}_0^s) \sum_{\tilde{W}\subseteq \alpha(\tilde{v}_0)} \Bigg(\prod_{w\in W} P_\theta(w)\Bigg) \Bigg[P\Bigg(G-\{v_0,v'_0\}-W, \bigcap_{\tilde{u}\in(U\backslash\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\backslash\tilde{W})} \tilde{u}^{y(\tilde{u})}\Bigg)
\end{aligned}
\tag{2}
$$

$$+P\left(G-\{v_0,v_0'\}-W, \bigcap_{\tilde{u}\in(U\setminus\tilde{v}_0)\cup(\alpha(\tilde{v}_0)\setminus\tilde{W})} \tilde{u}^{\bar{y}(\tilde{u})}\right)\Bigg]$$

$$=P\left(G, \bigcap_{\tilde{u}\in U} \tilde{u}^{x(\tilde{u})}\right)+P\left(G, \bigcap_{\tilde{u}\in U} \tilde{u}^{\bar{x}(\tilde{u})}\right)$$

where the inequality is obtained by using the induction hypothesis (i) and (ii).

Case 2. $U = \emptyset$ (both (i) and (ii) are reduced to $P_\theta(G) \geqq P(G)$). Choose any $\tilde{v}_0 = \{v_0, v_0'\} \in V_\theta$. Suppose that $P_\theta(v_0) \geqq P_\theta(v_0')$. Define

$$\beta(\tilde{v}_0)=\begin{cases}0 & \text{if } [v_0,v_0']\in E,\\ 1 & \text{if not.}\end{cases}$$

Then

$$P_\theta(G)=[P_\theta(v_0)P_\theta(v_0')]P_\theta(G-\{v_0,v_0'\})+[(1-P_\theta(v_0))(1-P_\theta(v_0'))]$$

$$\cdot\beta(\tilde{v}_0)\left[\prod_{v\in d(\tilde{v}_0)} P_\theta(v)\right]P_\theta(G-\{v_0,v_0'\}-d(v_0))+P_\theta(G,\tilde{v}_0^l)+P_\theta(G,\tilde{v}_0^s)$$

$$\geqq[P(v_0)P(v_0')]P(G-\{v_0,v_0'\})+(1-P(v_0))(1-P(v_0')) \tag{3}$$

$$\cdot\beta(\tilde{v}_0)\left[\prod_{v\in\{\alpha(\tilde{v}_0)\}} P(v)\right]P(G-\{v_0,v_0'\}-d(v_0))+P(G,\tilde{v}_0^l)+P(G,\tilde{v}_0^s)$$

$$=P(G)$$

where the inequality is obtained by using the induction hypothesis (ii) and inequality (2). □

COROLLARY. *The inequalities in Theorem* 2 *are strict if* $P_\theta \neq P$ *and* $p_1 > 0$.

Proof. We prove this by induction on $|V_\theta|$. Note that if $P_\theta \neq P$, G_θ must have a singular edge for P, so $|V_\theta| \geqq 2$.

Let $V = \{\tilde{v}_1, \tilde{v}_2\}$. Since $[\tilde{v}_1, \tilde{v}_2]$ is singular for P, we can assume that $P(v_1) > P(\theta(v_1))$ and that $P(v_2) > P(\theta(v_2))$ without loss of generality. Then G has four possibilities as shown in Fig. 1. It is not hard to verify the corollary directly for all cases.

Next, consider the case $|V_\theta| \geqq 3$. Choose $\tilde{v}_0$ such that $G - \{v_0, v_0'\}$ retains a singular edge. Then

$$P_0(G-\{v_0,v_0'\})>P(G-\{v_0,v_0'\})$$

by induction. Furthermore, $P_\theta(v_0)P_\theta(v_0') = P(v_0)P(v_0') > 0$. Hence inequality (3) is strict. □

3. Applications. We give some examples of how the theorems in § 2 can be used to determine optimal assignments for consecutive-2 graphs.

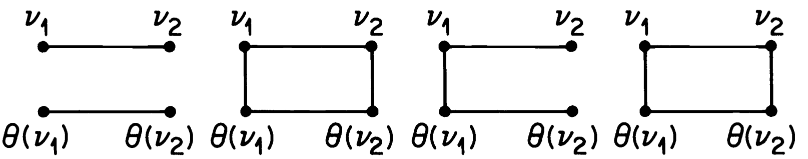


FIG. 1. *Four cases for* $|V_\theta| = 2$.

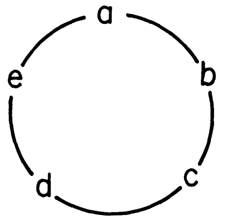


FIG. 2. *A 5-cycle.*

Example 1. Consider a 5-cycle as shown in Fig. 2 and a set of probabilities $p_1 < p_2 < p_3 < p_4 < p_5$.

By symmetry we may assume without loss of generality that p_5 is assigned to a. Consider the regular automorphism:

$$\theta(a, b)\colon a \leftrightarrow b, c \leftrightarrow e, d \leftrightarrow d.$$

By Theorem 1 there exists an assignment $P_{\theta(a,b)}$ without a singular edge with respect to $\theta(a, b)$ which satisfies

$$P_{\theta(a,b)}(c) > P_{\theta(a,b)}(e).$$

By the corollary of Theorem 2 any assignment violating the above inequality can be improved in the working probability for the graph. We denote this fact by the notation $c > e$ which means that the probability assigned to c should be larger than the one assigned to e.

Similarly, by considering the regular automorphisms

$$\theta(a, c)\colon a \leftrightarrow c, e \leftrightarrow d, b \leftrightarrow b,$$

$$\theta(a, d)\colon a \leftrightarrow d, b \leftrightarrow c, e \leftrightarrow e,$$

$$\theta(a, e)\colon a \leftrightarrow e, b \leftrightarrow d, c \leftrightarrow c, \text{ and}$$

$$\theta(a, a)\colon a \leftrightarrow a, b \leftrightarrow e, c \leftrightarrow d,$$

we obtain $d > e$, $c > b$, $d > b$ and $b > e$, $d > c$. By symmetry we may assume without loss of generality that the probability assigned to b is larger than the one assigned to e. Thus we obtain the ordering $a > d > c > b > e$, i.e., an optimal assignment is $p_5 \rightarrow a$, $p_2 \rightarrow b$, $p_3 \rightarrow c$, $p_4 \rightarrow d$ and $p_1 \rightarrow e$.

An analogous argument allows us to obtain an optimal assignment for any n-cycle.

Example 2. Let G consist of two 4-lines as shown in Fig. 3 and

$$p = \{p_1 < p_2 < \cdots < p_8\}.$$

Without loss of generality, assume that $a > e > h$ and $a > d$. Consider the regular automorphisms:

$$\theta(a, e)\colon a \leftrightarrow e, b \leftrightarrow f, c \leftrightarrow g, d \leftrightarrow h,$$

$$\theta(a, h)\colon a \leftrightarrow h, b \leftrightarrow g, c \leftrightarrow f, d \leftrightarrow e,$$

$$\theta(a, d)\colon a \leftrightarrow d, b \leftrightarrow c, e \leftrightarrow h, f \leftrightarrow g.$$

$$a - b - c - d$$

$$e - f - g - h$$

FIG. 3. *Two 4-lines.*

$$p_4 - p_5 - p_8 - p_1$$

$$p_3 - p_6 - p_7 - p_2$$

FIG. 4. *Optimal assignment for two* 4*-lines.*

We obtain

$$a > e, \quad b > f, \quad c > g, \quad h > d,$$
$$a > h, \quad g > b, \quad c > f, \quad e > d,$$
$$a > d, \quad c > b, \quad e > h, \quad g > f.$$

Thus we obtain the partial order

$$a > e > h > d \quad \text{and} \quad c > g > f > b.$$

However, a 4-line assignment problem is equivalent to a 5-cycle assignment problem with one additional probability 1 (since we can cut open the cycle at the vertex assigned with probability 1). From Example 1, an optimal 4-line with $p' = \{p_1' < p_2' < p_3' < p_4'\}$ should be

$$p_2' - p_3' - p_4' - p_1'.$$

Therefore $b > a$ and we obtain the linear order

$$c > g > f > b > a > e > h > d.$$

An optimal assignment is as shown in Fig. 4.

An analogous argument allows us to obtain an assignment for any two n-lines for even n. For odd n θ is not regular since it violates condition (iii), so Theorems 1 and 2 do not apply.

Example 3. A (k, m) caterpillar is a graph containing a path of m vertices of degree $k \geqq 2$ and having all other vertices of degree 1. Theorem 1 and the corollary of Theorem 2 can be used to obtain optimal assignments for arbitrary (k, m) caterpillars. Here we only illustrate an optimal assignment for the (3, 4) caterpillar as shown in Fig. 5 with $p = \{p_1 < p_2 < \cdots < p_{10}\}$.

Without loss of generality, assume that $a > i > j$ and $a > b$. Consider the regular automorphisms:

$$\theta(a, i): a \leftrightarrow i, b \leftrightarrow j, c \leftrightarrow h, d \leftrightarrow f, e \leftrightarrow g,$$

$$\theta(a, j): a \leftrightarrow j, b \leftrightarrow i, c \leftrightarrow h, d \leftrightarrow f, e \leftrightarrow g.$$

We obtain $a > b > i > j$, $h > c$, $d > f$, $g > e$.

Let G' be the graph obtained from G by adding edges $[b, x]$, $[b, y]$, $[y, z]$ and $[y, w]$ as shown in Fig. 6.

If we assign probability 1 to x and y, probability $p_1/2$ to z and w, then any assignment on G corresponds to an assignment on G' with the same graph working probability.

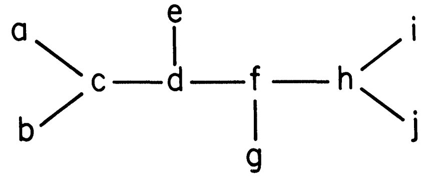


FIG. 5. *A* (3, 4) *caterpillar.*

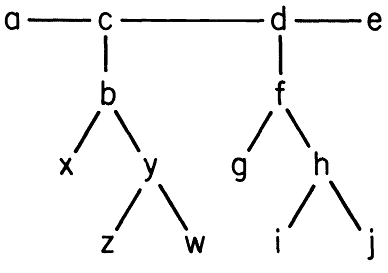


FIG. 6. *The graph G'.*

So an optimal assignment on G', subject to the restriction we specified, induces an optimal assignment on G. Consider the regular automorphism $\theta(a, e)$, $\theta(a, e)$: $a \leftrightarrow e$, $c \leftrightarrow d$, $b \leftrightarrow f$, $x \leftrightarrow g$, $y \leftrightarrow h$, $z \leftrightarrow i$, $w \leftrightarrow j$. Since $i > z$, $j > w$, $x > g$ and $y > h$, we obtain $f > b$, $c > d$, $e > a$.

Similarly, let G'' be the graph obtained from G by adding edges $[g, d']$, $[g, h']$, $[d', e']$, $[d', c']$, $[c', a']$, $[c', b']$, $[h', i']$ and $[h', j']$ as shown in Fig. 7.

Again, by assigning probability 1 to a', b', d', h' and probability $p_1/2$ to c', e', i', j', an optimal assignment on G'' induces an optimal assignment on G.

Consider the regular automorphism $\theta(f, g)$ on G'', $\theta(f, g)$: $f \leftrightarrow g$, $v \leftrightarrow v'$ for $v \in \{a, b, c, d, e, h, i, j\}$. Since $d' > d$ and $h' > h$, we obtain $f > g$.

Combining all the paired comparisons of vertices, we obtain the linear order $h > c > d > f > g > e > a > b > i > j$, i.e., an optimal assignment of the (3, 4) caterpillar is as shown in Fig. 8.

Example 4. We give an optimal assignment of $p = \{p_1 < p_2 < \cdots < p_8\}$ for a cube in Fig. 9 (the proof of optimality is left to the reader).

4. A remark. In the last section, every example has an optimal assignment which depends only on the linear order of p_i's. We call such optimal assignments *invariant optimal assignments*. However, invariant optimal assignments do not exist in general (see [3] for examples). An interesting question is what graph G has an invariant optimal assignment.

To see a sufficient and necessary condition, we first formalize the technique we used in Example 3, which is to add some vertices with working probabilities 1 or $\frac{1}{2}p_1$. Let $G = (V, E)$ be a subgraph of $G' = (V', E')$. We call G' a *feasible extension* of G if we can

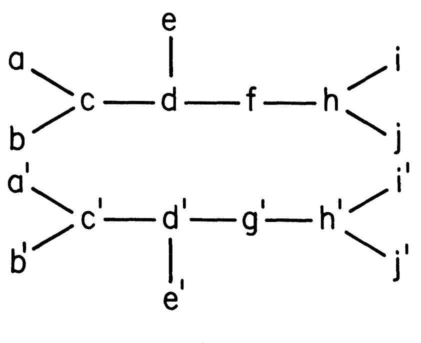


FIG. 7. *The graph G''.*

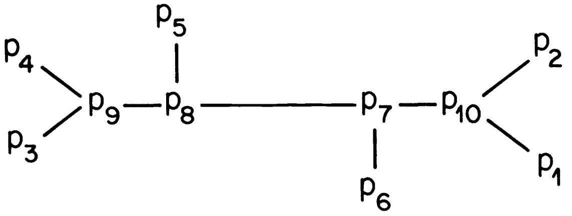


FIG. 8. *An optimal assignment for the* (3, 4) *caterpillar.*

assign 1 and $\frac{1}{2}p_1$ to $G' - G$ such that:

(1) $[u, v] \in E'$, $u \in V$, $v \in V' - V$ imply $P(v) = 1$;

(2) $[u, v] \in E'$, $u, v \in V' - V$ imply $P(u) = 1$, $P(v) = \frac{1}{2}p_1$ or $P(u) = \frac{1}{2}p_1$, $P(v) = 1$.

A regular automorphism θ of G' is *feasible* for G if, for any assignment P to G' satisfying (1) and (2), there exists an assignment P_θ to G' such that

(i) $\{P_\theta(v), P_\theta(\theta(v))\} = \{P(v), P(\theta(v))\}$;

(ii) G' has no singular edge with respect to P_θ;

(iii) $u \in V' - V$ implies $P_\theta(u) = P(u)$.

By an argument similar to the proof of Theorem 1, we see that θ is feasible for G if and only if the following condition holds. Let u', $v' \in V' - V$, u, $v \in V$ and $[u, u'] \in E'$, $[v, v'] \in E'$. Then for any path $(\bar{u}, \bar{u}_0, \cdots, \bar{u}_m, \bar{v}')$ in G'_θ,

$$|\{[u', u_0], [u_0, u_1], \cdots, [u_m, v']\} \cap E|$$

is always an even number.

Let u, $v \in V$. We say that u, v are *comparable if there exists a feasible extension G' of G and a feasible regular automorphism θ of G' for G such that $\theta(u) = v$.*

THEOREM 3. *If all pairs u, v of V are comparable, then G has an invariant optimal assignment.*

Proof. Let u, $v \in V$ be compared under G' and θ. Suppose that there is a $w \in V' - V$, which is adjacent to a vertex of G, such that G' has a path ξ from u to w. For any optimal assignment P of G, we extend P to G' such that (1) and (2) hold. Then G'_θ has no singular edge under P. Note that $P(w) = 1 > P(\theta(w))$. Thus, $P(u) > P(v)$ if the path ξ contains an even number of edges and $P(u) < P(v)$ if the path ξ contains an odd number of edges. For this reason, we define $v < u$ if the former case occurs and $u < v$ if the latter case occurs. The relation "<" induces a partial ordering on V. We first prove that for any minimal element u, there exists an optimal assignment P^* to G such that $P^*(u) = p_1$. Suppose that P is an optimal assignment to G and $P(v) = p_1$. It is easy

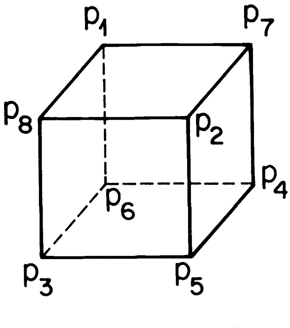


FIG. 9. *An optimal assignment for a* 3-*cube.*

to see that v is a minimal element of V. Consider G' and θ under which u and v are compared. Then, there is no $w \in V' - V$ such that G' has a path from w to u or v. Let C be a connected component of G'_θ containing $\{u, v\}$. Define $P^*(x) = P(\theta(x))$ if $\tilde{x}$ is a vertex of C and $P^*(x) = P(x)$ if $\tilde{x}$ is not a vertex of C. Then G'_θ has no singular edge under P^* also. Therefore, $P(G) = P(G') = P_\theta(G') = P_\theta(G)$, that is, P^* is an optimal assignment to G such that $P^*(u) = p_1$.

Suppose that $p_1, \cdots, p_k$ have been assigned to $v_1, \cdots, v_k$ $(k \geqq 1)$. We now assign p_{k+1}. Extend the ordering "<" on V by considering every two vertices, u, v of V. If u and v are not ordered under the ordering "<," we find G' and θ under which u and v are compared. If G' has a path from v_k to u or v, say u, and $\theta(v_k) \neq v_k$, then define $v < u$ if this path contains an even number of edges, and $u < v$ if this path contains an odd number of edges. After extending this ordering, assign p_{k+1} to a minimal element v_{k+1} of $V - \{v_1, \cdots, v_k\}$ under this ordering. We need to prove that there exists an optimal assignment P^* to G such that $P^*(v_i) = p_i$, $i = 1, \cdots, k+1$. Note that by the induction hypothesis, there is an optimal assignment P to G such that $P(v_i) = p_i$, $i = 1, \cdots, k$. Suppose $P(u) = p_{k+1}$ and $u \neq v_{k+1}$. It is easy to see that u is also a minimal element of $V - \{v_1, \cdots, v_k\}$ under the extended ordering "<". Thus, u and v_{k+1} are not ordered under the extended ordering "<". Let u and v_{k+1} be compared under G' and θ. Let C be a connected component of G'_θ containing $\{u, v_{k+1}\}$. We can show that for any $w \in V' - V$, we have $\tilde{w} \notin C$ and if $\tilde{v}_i \in C$ for some $i = 1, \cdots, k$ then $v_i = \theta(v_i)$. Therefore, we can obtain the required optimal assignment P^* to G by defining $P^*(x) = P(x)$ if $\tilde{x} \notin C$ and $P^*(x) = P(\theta(x))$ if $\tilde{x} \in C$. □

Conjecture. If G has an invariant optimal assignment, then all pairs u, v of V are comparable.

REFERENCES

[1] D. DERMAN, G. J. LIEBERMAN AND S. M. ROSS, *On the consecutive-k-out-of-n system*, IEEE Trans. Reliability, 31 (1982), pp. 57–63.

[2] D. Z. DU AND F. K. HWANG, *Optimal consecutive-2-out-of-n system*, Math. Oper. Res., 11 (1986), pp. 187–191.

[3] ———, *Optimal consecutive-2 systems of lines and cycles*, Networks, 15 (1985), pp. 439–447.

[4] V. K. WEI, F. K. HWANG AND V. T. SÓS, *Optimal sequencing of items in a consecutive-2-out-of-n system*, IEEE Trans. Reliability, 32 (1983), pp. 30–33.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

001

SS/TDMA SATELLITE COMMUNICATIONS WITH *k*-PERMUTATION SWITCHING MODES*

J. L. LEWANDOWSKI† AND C. L. LIU‡

Abstract. The Satellite-Switched Time-Division Multiple Access (SS/TDMA) scheme has been one of the most effective techniques designed to allocate the communication bandwidth provided by communication satellites. The scheduling problem for SS/TDMA corresponds to finding a positive linear combination of a predefined set of (0, 1)-matrices which covers a given traffic matrix **T** such that the sum of the multiplying constants used in the linear combination is minimum. In this paper, an algorithm is given to solve the optimization problem using a result which is a generalization of a theorem by Birkhoff and von Neumann. The case of *k*-permutation matrices is first addressed. The result is then further extended to more general sets of (0, 1)-matrices.

Key words. Birkhoff–von Neumann, network flow, combinatorial optimization

AMS(MOS) subject classifications. 08-04, 90C27

1. Introduction. In recent years, many wideband satellite communication systems have been constructed to link together a large number of earth stations. The Satellite-Switched Time-Division Multiple Access (SS/TDMA) method is one of the most effective techniques designed to allocate the communication bandwidth provided by a satellite link to carry the traffic between earth stations [5]. In an SS/TDMA system, a number of spot-beam antennas, on-board the satellite, divides the coverage area into spatially disjoint common access channels. Traffic from an earth station is sent via an up-link beam to the satellite and is routed by an on-board processor to a down-link beam which is received by another earth station. An on-board switch connection which specifies the interconnections of up-link beams to down-link beams via transponders is referred to as a *switching mode.*

In a typical TDMA implementation earth stations have different amounts of information to be transmitted to other earth stations. In general, there are a certain number of up-link and down-link beams, denoted by n. The satellite also has some processing capability which is used to route the information from the up-link to the down-link beams. Let t_{ij} be the amount of information which is to be routed from the uplink beam i to the downlink beam j. We refer to t_{ij} as the amount of traffic from i to j. This $n \times n$ matrix **T** is referred to as the *traffic matrix.* A particular choice of traffic matrix is referred to as a *TDMA frame.* To satisfy the traffic demand, each TDMA frame is divided into a number of time slots. In each of these time slots, a different switching mode is used so that traffic from earth stations in different beam zones can be routed to their destinations. More specifically, each switching mode is specified by an $n \times n$ (0, 1)-matrix which must satisfy certain technologically dependent constraints and which shall be referred to as a *switching mode.* We say that matrix **X** *dominates* matrix **Y** (which we denote by $\mathbf{X} \geqq \mathbf{Y}$) if each of the entries in **X** is at least as large as the corresponding entry in **Y**. Let $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_\gamma$ denote the switching modes used in a frame, and let c_i, $1 \leqq i \leqq \gamma$ be the length of time that the ith switching mode is used. The total amount of time available in the frame to route traffic between different beam zones is given by the

* Received by the editors February 22, 1983; accepted for publication (in revised form) October 14, 1986. This work was partially supported by the Office of Naval Research under grant N00014-86K-0416.

† AT&T Engineering Research Center, Princeton, New Jersey 08540.

‡ University of Illinois, Department of Computer Science, Urbana, Illinois 61801.

$n \times n$ matrix $\mathbf{T}'$:

$$\mathbf{T}' = \sum_{i=1}^{\gamma} c_i \mathbf{Z}_i.$$

Clearly, in order for all the traffic specified by $\mathbf{T}$ to be routed, $\mathbf{T}'$ must dominate $\mathbf{T}$. For a given traffic matrix $\mathbf{T}$, the problem of choosing a set of switching modes $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_\gamma$ in order to maximize the utilization of on-board transponders (i.e., to minimize the total amount of transmission time) $\sum_{i=1}^{\gamma} c_i$, which is known as the *cost* of the decomposition, is referred to as the time-slot decomposition problem.

In § 2, we describe the current problem and present the definitions needed in this paper. In § 3, an algorithm for solving this problem is presented, and its running time is analyzed. In § 4, a result is given on the performance of the decomposition generated by our algorithm. In § 5, a generalization of this algorithm is presented.

2. Definitions. In general, each up-link and down-link beam is a multiplex of several different signals. There are certain constraints, imposed by hardware considerations, on the amount of demultiplexing of up-link beams, the ability to switch these signals, and the amount of multiplexing that can be done into the down-link beams. We consider the case where each of the up-link and down-link beams is a multiplex of no more than a certain number k of signals. We therefore assume that the satellite is able to demultiplex each of the up-link beams into k signals, switch each of these signals to different down-link beams, and multiplex up to k of these signals into each of the down-link beams. That is, the switching modes to be used are (0, 1)-matrices in which there are no more than k 1's in each row and column. Mathematically, the problem can be formulated as follows: Let $\mathbf{V}(\mathbf{k})$ denote the set of all (0, 1)-matrices which have no more than k 1's in each row, and no more than k 1's in each column. Given a matrix $\mathbf{T}$ of nonnegative entries, we wish to find positive constant $\nu_1, \nu_2, \cdots, \nu_\gamma$ and matrices $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_\gamma$ in $\mathbf{V}(\mathbf{k})$ such that

$$\mathbf{T} \leqq \mathbf{T}' = \sum_{i=1}^{\gamma} \nu_i \mathbf{Z}_i \tag{2.1}$$

with $\sum_{i=1}^{\gamma} \nu_i$ being minimum. In this paper, we solve a special case of this problem by restricting ourselves to the class of (0, 1)-matrices which have *exactly* k 1's in each row, and *exactly* k 1's in each column. We let $\mathbf{U}(\mathbf{k})$ denote this class of matrices, and call its members *k-permutation switching modes*. In this paper, we solve the time slot assignment problem for switching modes $\mathbf{U}(\mathbf{k})$, and give a bound on the performance of this algorithm with respect to the optimal schedule for switching modes in $\mathbf{V}(\mathbf{k})$. This, and the generalization presented in § 5, solve a problem related to an open problem mentioned in [2].

3. Construction of T′. We state without proof the following theorem [3], which is a generalization of the celebrated Birkhoff–von Neumann Theorem.

THEOREM 3.1. *Let* $\mathbf{T}'$ *be a matrix with nonnegative entries. The following conditions are necessary and sufficient for* $\mathbf{T}'$ *to be expressible as a convex sum of* (0, 1)*-matrices in* $\mathbf{U}(\mathbf{k})$.

(i) $r_i' = kx$ *for all* i,

(ii) $c_j' = kx$ *for all* j, *and*

(iii) $t_{ij}' \leqq x$ *for all* i *and* j

where r_i' *denotes the sum of the entries in the ith row of* $\mathbf{T}'$, *and* c_j' *denotes the sum of the entries in the jth column of* $\mathbf{T}'$, *and* x *is a constant.*

It follows immediately that if $\mathbf{T}' = \sum_{i=1}^{\gamma} v_i \mathbf{Z}_i$, then $\sum_{i=1}^{\gamma} v_i = x$. Thus, our problem is reduced to finding a matrix $\mathbf{T}'$ such that (a) $\mathbf{T}' \geqq \mathbf{T}$, (b) $\mathbf{T}'$ satisfies the three conditions of Theorem 3.1, and (c) x is as small as possible. From (iii) is follows that x must be larger than or equal to the largest entry of $\mathbf{T}$. From (i) it is seen that x must also be larger than or equal to the ith row sum r_i divided by k for every i. Similarly, from (ii) it must also be larger than or equal to the jth column sum c_j divided by k for every j. We therefore have an initial lower bound on the value x:

$$x = \max_{i,j} \left(t_{ij}, \frac{r_i}{k}, \frac{c_j}{k} \right).$$

Construct the network in Fig. 1. There are n edges from the source a, which correspond to the n row sums of $\mathbf{T}$, with the capacity of the ith edge being the total amount that the entries in the ith row must be increased in order to satisfy condition (i) in Theorem 3.1, namely: $kx - r_i$. Similarly, there are n edges to the sink z, which correspond to the n column sums of $\mathbf{T}$ with the capacity of the jth edge being the total amount that the entries in the jth column must be increased in order to satisfy condition (ii), namely: $kx - c_j$. There are also n^2 interior edges which correspond to the n^2 entries of $\mathbf{T}$. The edge from the ith node on the left-hand side to the jth node on the right-hand side corresponds to the ijth entry of $\mathbf{T}$, and its capacity is the largest amount that can be added to this entry without condition (iii) being violated; namely: $x - t_{ij}$. By our choice of x, it is easy to see that all of these capacities are nonnegative, so a maximum flow (minimum cut) can be found by well-known methods [1].

A flow which saturates all of the edges coming out of the source for some value x will be called a *row saturating flow*. The value of such a flow is given by $\sum_{i=1}^{n} (kx - r_i)$. It is immediate that such a flow will also saturate all of the edges going into the sink, which we recall correspond to the column sums. We have the following result.

THEOREM 3.2. *There is a* 1-1 *correspondence between row saturating flows in this network, and matrices* $\mathbf{T}'$ *which satisfy the three conditions of Theorem* 3.1.

Proof. Given a row saturating flow, if each entry in $\mathbf{T}$ is increased by the amount of the flow in the corresponding edge in the network, the resulting matrix $\mathbf{T}'$ clearly satisfies the three conditions of Theorem 3.1. Conversely, given a matrix $\mathbf{T}'$ which satis-

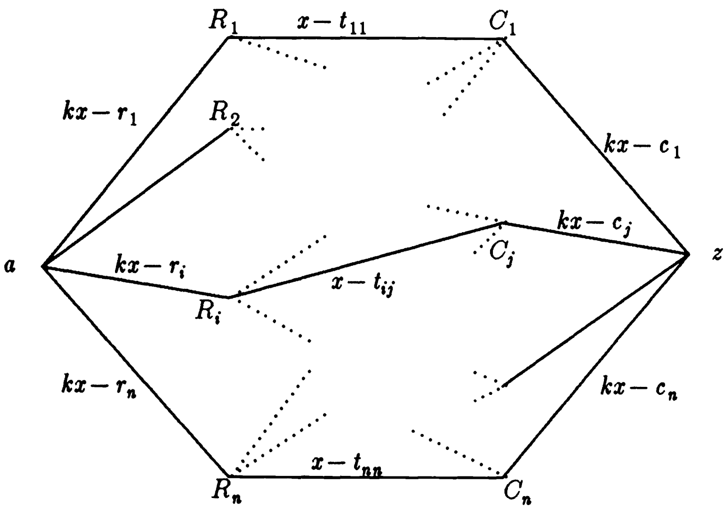


FIG. 1. *The network.*

fies these conditions, let the flow through the edge connecting the ith node on the left to the jth node on the right be $t'_{ij} - t_{ij}$, the flow through the ith edge on the left-hand side be $\sum_{j=1}^{n} (t'_{ij} - t_{ij})$, and the flow through the jth edge on the right-hand side be $\sum_{i=1}^{n} (t'_{ij} - t_{ij})$. Clearly, the resulting flow is row saturating. □

Therefore, we proceed by using the initial bound on x and try to find a flow which is row saturating. If one can be found, $\mathbf{T}'$ can be constructed. However, if no such flow can be found, x must be increased, and the procedure iterated.

Any cut in this network can be put into the standard form shown in Fig. 2. This is done by permuting the edges of the left- and right-hand sides so that the first t edges on the left-hand side are not in the cut, and the first s edges on the right-hand side are in the cut. (This corresponds to permuting rows and columns of $\mathbf{T}$ and the corresponding rows and columns of all the switching modes in $\mathbf{U(k)}$, which can be done without loss of generality.) The general form for the capacity of such a cut is given by

$$\sum_{i=t+1}^{n} (kx - r_i) + \sum_{j=1}^{s} (kx - c_j) + \sum_{i=1}^{t} \sum_{j=s+1}^{n} (x - t_{ij}).$$

If a cut F whose capacity is less than that of a row saturating flow is found, x must be increased by some positive value δ so that the capacity of F will be at least as large as a row saturating flow. If we replace x by $x + \delta$, the following inequality must hold:

$$\sum_{i=t+1}^{n} [k(x+\delta) - r_i] + \sum_{j=1}^{s} [k(x+\delta) - c_j] + \sum_{i=1}^{t} \sum_{j=s+1}^{n} [(x+\delta) - t_{ij}] \geqq \sum_{i=1}^{n} [k(x+\delta) - r_i]$$

which can be written as

$$\left[\sum_{i=t+1}^{n} k + \sum_{j=1}^{s} k + (n-s)t\right]\delta$$
$$+\left[\sum_{i=t+1}^{n} k + \sum_{j=1}^{s} k + (n-s)t\right]x - \left[\sum_{i=t+1}^{n} r_i + \sum_{j=1}^{s} c_j + \sum_{i=1}^{t} \sum_{j=s+1}^{n} t_{ij}\right]$$
$$\geqq \sum_{i=1}^{n} k\delta + \sum_{i=1}^{n} (kx - r_i).$$

If we let

$$A = \sum_{i=1}^{n} k,$$

$$B = \sum_{i=1}^{n} (kx - r_i),$$

$$C = \sum_{i=t+1}^{n} k + \sum_{j=1}^{s} k + (n-s)t,$$

$$D = Cx - \left[\sum_{i=t+1}^{n} r_i + \sum_{j=1}^{s} c_j + \sum_{i=1}^{t} \sum_{j=s+1}^{n} t_{ij}\right],$$

the following must hold:

$$C\delta + D \geqq A\delta + B.$$

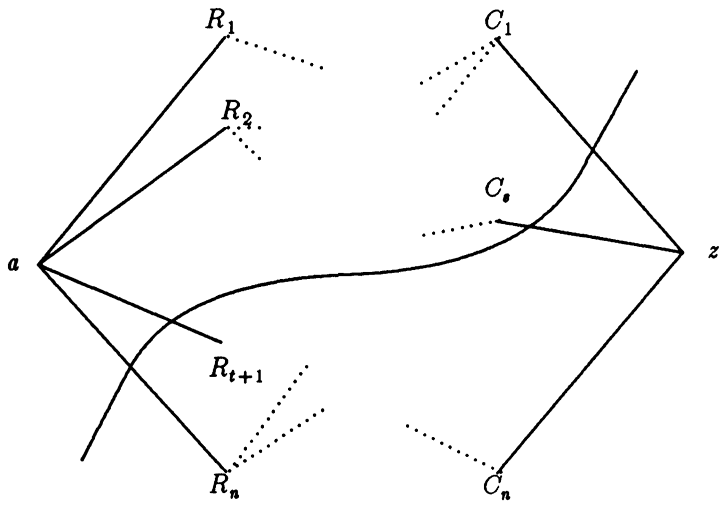


FIG. 2. *The standard form for a cut.*

By the assumption that the flow is not row saturating, $B > D$. In order to show that δ can be increased in order to make the capacity of the cut which has been found smaller than a the cut F, it must be shown that $C > A$. This is equivalent to showing that

$$(s-t)k+(n-s)t>0 \tag{3.1}$$

or, equivalently

$$sk+nt>kt+st. \tag{3.2}$$

Since $n > s$ and $t > 0$ for nonrow saturating flows, $(n - s)t > 0$, and (1) holds if $s \geqq t$. We can assume, therefore, that $t > s$. Since $nt > kt$, (2) holds if $k \geqq t$; therefore we assume that $t > k$. Thus, (1) holds if $n - s \geqq s - t$, or $n + t \geqq 2s$. But we have assumed that $t > s$, so (1) holds. From this argument it is seen that the smallest possible δ which can be chosen is $\delta = (B - D)/(C - A)$.

3.1. Algorithm Expand. In this section, we present an algorithm for expanding the entries of an $n \times n$ matrix $\mathbf{T}$ so that the resulting matrix $\mathbf{T}'$ satisfies the three requirements of Theorem 3.1.

```
ALGORITHM EXPAND.
Step 0:
    Let $x = \max(\max_{i,j}(t_{ij}), \max_i(r_i/k), \max_j(c_j/k))$.
Step 1:
    Construct the network described above, find a maximum flow (minimum cut), and
    let $F$ denote its capacity.
Step 2:
    If $F = \sum_{i=1}^{m}(kx - r_i)$, a row saturating flow has been found, go to Step 5.
Step 3:
    Compute the capacities of this minimum cut and of the row saturating flow in the
    form $C\delta + D$, and $A\delta + B$, respectively, as above.
```

Step 4:

Let $\delta = (B - D)/(C - A)$, which is a positive value by our previous argument, $x = x + \delta$, and return to Step 1.

Step 5:

Increase each of the entries in **T** by the amount of the flow through the corresponding interior edges in this network, forming the new matrix **T**′, which at this point satisfies the conditions of Theorem 3.1. **HALT.**

The correctness of the algorithm follows from the above discussion. Termination is insured by the fact that there are only finitely many cuts in the network, while each iteration of Steps 1–4 satisfies the capacity requirement of at least one additional cut. (However, for a better estimate of the running time of Algorithm Expand, see § 3.2.) The algorithm given in [3] can now be used to decompose **T**′ into a sum of switching modes in **U(k)**. The minimality of the sum of the coefficients ν_i in equation (2.1) is guaranteed by the fact that at each iteration of Step 4, x is incremented by the smallest amount which might allow for the expansion to be possible. As an example of the expansion of a traffic matrix **T** using Algorithm Expand, we consider the traffic matrix in Fig. 3. The initial bound for x is 10. The corresponding network is shown in Fig. 4(a), and the maximum flow (whose value is 0) is shown in Fig. 4(b) with the edges in the corresponding cut marked. This cut is of the form 13δ, and a row saturating cut is of the form $8\delta + 1$. Therefore we select .2 as our δ and let 10.2 be our new value for x. Figure 4(c) shows the new network corresponding to this value for x. Figure 4(d) shows the maximum flow in this network, which in this case is row saturating. Finally, Fig. 4(e) shows the corresponding traffic matrix **T**′ which satisfies the three requirements of Theorem 3.1.

3.2. Analysis of the running time of Algorithm Expand. In this section, the running time of Algorithm Expand is analyzed. We first show that the coefficients of δ which occur during the execution of Algorithm Expand lie in specific range. Using this fact, the number of iterations of the algorithm is bounded.

By examination, the coefficient of δ (C in our notation) is at least as large as $\sum_{i=1}^{n} k$ for every cut in the network. It can also be seen that this coefficient must be integral. It is, however, no larger than

$$\sum_{i=1}^{n} k + \sum_{j=1}^{n} k + nt$$

or

$$2nk + nt.$$

This value may not be attained for any cut in the network, but is useful for the analysis. This value is on the order of $O(n^2)$, and in fact can be no larger than $3n^2$. Therefore, C can only take on values between n and $3n^2$. Arrange the capacities of all of the cuts of

$$\mathbf{T} = \begin{bmatrix} 1 & 6 & 2 & 10 \\ 8 & 1 & 10 & 1 \\ 8 & 8 & 1 & 3 \\ 3 & 5 & 7 & 5 \end{bmatrix}$$

FIG. 3. *The traffic matrix to be expanded.*

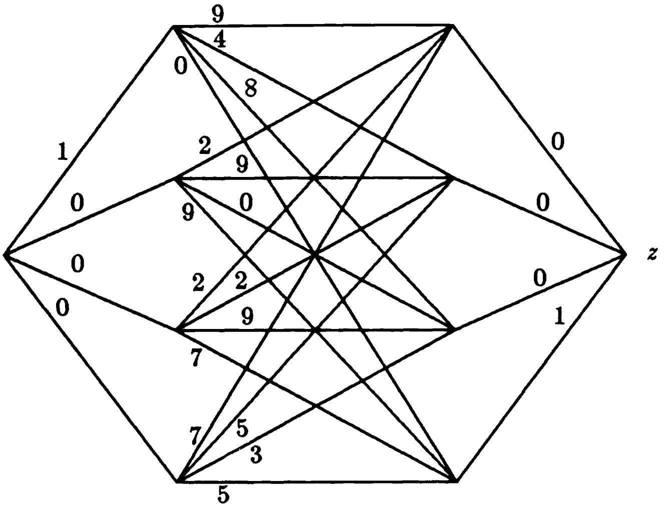


FIG. 4(a). *Network and capacities for* $x = 10$.

the network in the form $C_i\delta + D_i$ in increasing lexicographical order. That is, $C_i\delta + D_i \geqq C_j\delta + D_j$ if $C_i > C_j$, or $C_i = C_j$ and $D_i \geqq D_j$. If $C_i\delta + D_i$ is at least as large as a row saturating flow, then it is easy to show that $C_j\delta + D_j$ is at least as large as a row saturating flow for all $C_j\delta + D_j$ which are lexicographically larger than $C_i\delta + D_j$. This proves that each coefficient of δ appears at most once during the execution of Algorithm Expand; and therefore, the algorithm terminates after at most $O(n^2)$ iterations.

The time for each iteration of the algorithm is bounded by the construction of the network, and the calculation of the maximum flow. The construction can be done in $O(n^2)$ time, the maximum flow can be computed in time $O(n^3)$ [1]. Therefore, the total running time of this algorithm is $O(n^5)$.

4. Analysis of the decomposition generated by Algorithm Expand. A solution for the decomposition problem using switching modes in the class $\mathbf{U(k)}$ has been presented.

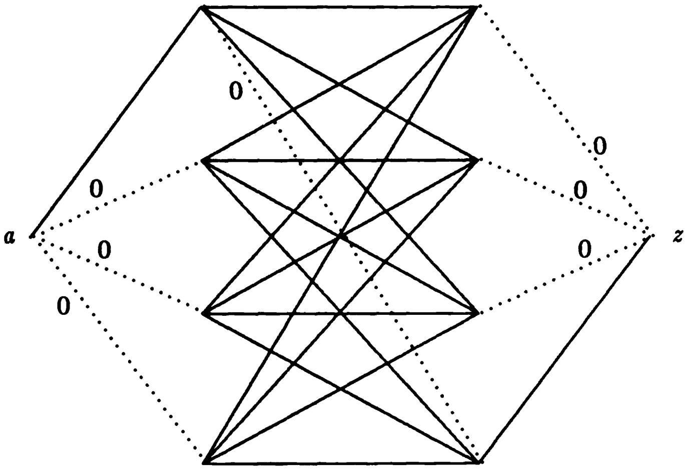


FIG. 4(b). *Maximum flow for network in Fig.* 4(a).

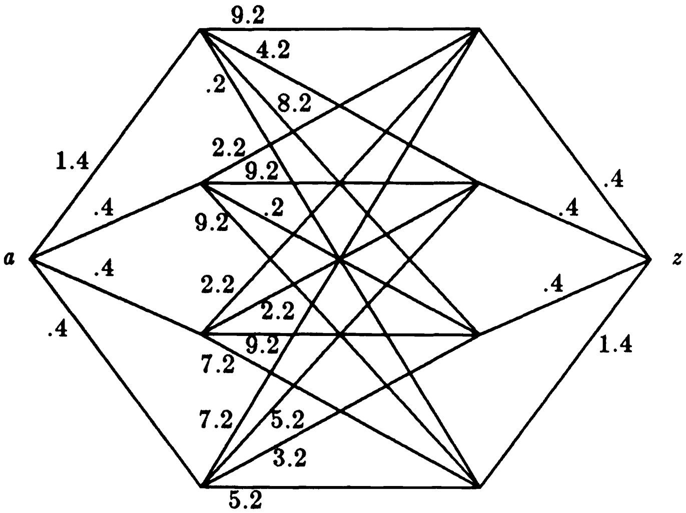


FIG. 4(c). *Network and capacities for* $x = 10.2$.

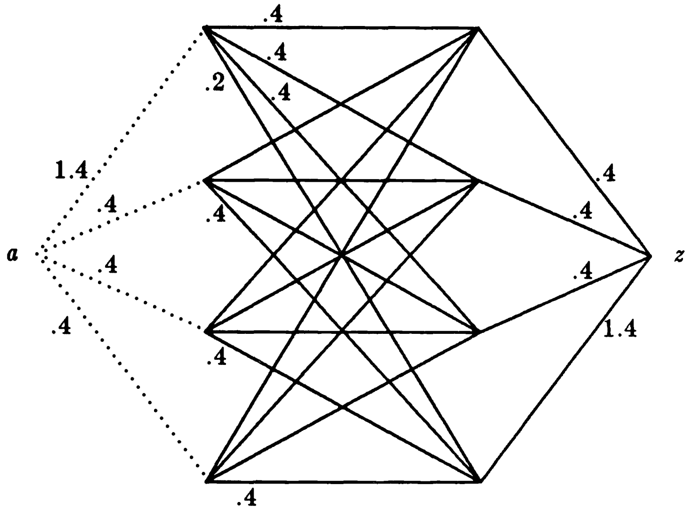


FIG. 4(d). *Maximum flow for network in Fig.* 4(c).

$$\mathbf{T}' = \begin{bmatrix} 1.4 & 6.4 & 2.4 & 10.2 \\ 8 & 1 & 10 & 1.4 \\ 8 & 8 & 1 & 3.4 \\ 3 & 5 & 7 & 5.4 \end{bmatrix}$$

FIG. 4(e). *Corresponding traffic matrix* T′.

However, the problem of decomposing a matrix into a convex sum of switching modes in $\mathbf{V(k)}$ is perhaps of more practical interest. It has been assumed until now that there are exactly k 1's in each row and column of the switching modes. But in practice, all that is required is that there are no more than k 1's in any row or column. That is, none of the technical limitations of the system should be exceeded. We should hope that forcing *exactly* k 1's in each row and *exactly* k 1's in each column does not cost very much. In this section, we prove that, indeed, the cost of the decomposition generated by Algorithm Expand for switching modes in $\mathbf{U(k)}$ is no worse than twice that of the optimal decomposition using switching modes in $\mathbf{V(k)}$.

As an example of a traffic matrix which is decomposable into switching modes in $\mathbf{V(k)}$ with unit cost, but which can not be decomposed into switching modes in $\mathbf{U(k)}$ with this cost, consider the one shown in Fig. 5. Clearly, this matrix can be dominated by one switching mode in $\mathbf{V}(3)$. However, it is easy to see that no switching mode in $\mathbf{U}(3)$ can dominate this matrix. This implies that more than unit time is needed for any decomposition of this matrix into switching modes in $\mathbf{U}(3)$. (In fact, it can be shown that 9/7 is the optimal cost for decomposing this matrix into switching modes in $\mathbf{U}(3)$.) We shall use the following result.

THEOREM 4.1. *Any switching mode in* $\mathbf{V(k)}$ *can be dominated by a matrix which is decomposable into a sum of switching modes in* $\mathbf{U(k)}$ *with a cost of at most two.*

Proof. We consider two cases.

If $k \geqq n/2$, we note that the matrix $\mathbf{J}$ (that is, the $n \times n$ matrix all of whose entries are one) dominates every matrix in $\mathbf{V(k)}$. However, by the result of [3], this matrix can be decomposed into a convex sum of switching modes in $\mathbf{U(k)}$ with cost n/k, and therefore, the result holds.

If $k < n/2$, we use a direct argument to dominate each switching mode in $\mathbf{V(k)}$ by a sum of at most two switching modes in $\mathbf{U(k)}$. Let $\mathbf{Z}$ be any switching mode in $\mathbf{V(k)}$. Let $\mathbf{Z}_1$ be the (0, 1)-matrix which is identical to $\mathbf{Z}$ for rows i, $1 \leqq i \leqq \lceil n/2 \rceil$, and is zero elsewhere. Similarly, let $\mathbf{Z}_2$ be the (0, 1)-matrix which is identical to $\mathbf{Z}$ for rows i, $\lceil n/2 \rceil + 1 \leqq i \leqq n$, and zero elsewhere. If it can be shown that $\mathbf{Z}_1$ and $\mathbf{Z}_2$ can have some of their zero entries charged to ones so that the resulting (0, 1)-matrices are in $\mathbf{U(k)}$, the result follows. We consider only the matrix $\mathbf{Z}_1$, as the argument for $\mathbf{Z}_2$ is similar.

We first argue that some of the zeros in the upper half of $\mathbf{Z}_1$ can be changed to ones so that the ith row sum is equal to k for i, $1 \leqq i \leqq \lceil n/2 \rceil$. The procedure used is "greedy" in that all of the zeros are scanned once in some order (say from left to right, top to bottom), with a zero in the ijth location changed to a one if both the sum of the entries in the ith row and the sum of the entries in the jth column are less than k. This algorithm is run on the first $\lceil n/2 \rceil$ rows and columns; let $\mathbf{Z}'_1$ denote the resulting matrix. We claim that when this procedure terminates, the first $\lceil n/2 \rceil$ row sums are k. If some row sum is less than k, there must be at least $n - k + 1$ zeros in this row. Since none of these entries were changed to one during the execution of the "greedy" algorithm, each of the columns

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

FIG. 5. *A traffic matrix.*

$$\mathbf{Z}_1 = \begin{bmatrix} 1&0&0&1&1&0\\ 0&1&1&0&1&0\\ 1&1&0&0&1&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0 \end{bmatrix}$$

FIG. 6. *A partial k-permutation switching mode.*

$$\mathbf{K} = \begin{bmatrix} 1&4&5\\ 2&3&5\\ 1&2&5\\ 0&0&0\\ 0&0&0\\ 0&0&0 \end{bmatrix}$$

$$a_1 = 2 \quad a_2 = 2 \quad a_3 = 1 \quad a_4 = 1 \quad a_5 = 3 \quad a_6 = 0$$

FIG. 7. *The corresponding matrix* **K**.

$$\mathbf{K} = \begin{bmatrix} 1&4&5\\ 2&3&5\\ 1&2&5\\ 1&3&6\\ 2&4&6\\ 3&4&6 \end{bmatrix}$$

FIG. 8. *The completed* **K**.

$$\mathbf{Z}_1 = \begin{bmatrix} 1&0&0&1&1&0\\ 0&1&1&0&1&0\\ 1&1&0&0&1&0\\ 1&0&1&0&0&1\\ 0&1&0&1&0&1\\ 0&0&1&1&0&1 \end{bmatrix}$$

FIG. 9. *The completed k-permutation switching mode.*

must have sum k. For this to be true, there must be at least $(n - k + 1)k$ ones in the matrix. But since $k < \lceil n/2 \rceil$, this quantity is greater than $(\lceil n/2 \rceil + 1)k$. As there are only $\lceil n/2 \rceil$ rows, there is a contradiction; therefore, the first $\lceil n/2 \rceil$ row sums are k.

In order to complete this construction some of the 0's in the bottom half of $\mathbf{Z}'_1$ must be changed to 1's so that the resulting matrix is a k-permutation switching mode. It is easily seen that a k-permutation switching mode is equivalent to an $n \times k$ matrix $\mathbf{K}$ such that: (a) The entries in $\mathbf{K}$ are integers in $[1, n]$. (b) No entry appears more than once in any row of $\mathbf{K}$. (c) Each entry appears exactly k times. If we construct the partial $\mathbf{K}$ for the matrix $\mathbf{Z}_1$, we see that condition (a) is satisfied for all entries in the first $\lceil n/2 \rceil$ rows (where the other entries are filled in with zeros), condition (b) also holds for the first $\lceil n/2 \rceil$ rows, and no entry occurs more than k times. Let a_1 denote the number of 1's in this matrix, a_2 the number of 2's, $\cdots$, a_n the number of n's. Therefore, in order to complete the construction, the remaining entries must be filled in with $k - a_1$ 1's, $k - a_2$ 2's, $\cdots$, $k - a_n$ n's satisfying conditions (b) and (c). Start in the first column at the $(\lceil n/2 \rceil + 1)$th entry and proceed by columns in the region which has not been filled in and insert $k - a_1$ 1's, $k - a_2$ 2's, $\cdots$, $k - a_n$ n's. It is immediate that conditions (a) and (c) are satisfied for the resulting matrix. In order to see that condition (b) is also satisfied, we argue by contradiction. Assume that there is a row for which some value, say i appears (at least) twice. Since the entries were filled in by columns, the only way for there to be two i's in any row is if they have "wrapped around" one of the columns. This implies that $k - a_i$ is greater than the height of a column. However, these columns have height $\lfloor n/2 \rfloor$. This implies that $k > \lceil n/2 \rceil$, which contradicts our assumption that $k < n/2$. Therefore, all three conditions are satisfied, and the corresponding switching mode dominates $\mathbf{Z}_1$. As an example of this procedure, consider the matrix $\mathbf{Z}_1$ in Fig. 6, where $n = 6$, and $k = 3$, on which the "greedy" procedure has already been run. The corresponding matrix $\mathbf{K}$ which has been partially filled in is shown in Fig. 7, along with the a_i's. The matrix $\mathbf{K}$, after it has been completed is shown in Fig. 8, and the corresponding k-permutation switching mode $\mathbf{Z}'_1$ is shown in Fig. 9. This, with a similar argument to "expand" $\mathbf{Z}_2$, completes the argument that every switching mode in $\mathbf{V}(\mathbf{k})$ can be dominated by a sum of switching modes in $\mathbf{U}(\mathbf{k})$ with cost no more than two.

If we now consider an optimal decomposition of some traffic matrix $\mathbf{T}$ using switching modes in $\mathbf{V}(\mathbf{k})$ of the form:

$$\mathbf{T} \leqq \sum_{i=1}^{\gamma} \nu_i \mathbf{Z}_i$$

where each $\mathbf{Z}_i$ is a member of $\mathbf{V}(\mathbf{k})$, we can partition the $\mathbf{Z}_i$'s into two sets: those that are in $\mathbf{U}(\mathbf{k})$, and those that are not. After reordering the sum, this expression can be written:

$$\mathbf{T} \leqq \sum_{i=1}^{\alpha} \nu_i \mathbf{Z}_i + \sum_{i=\alpha+1}^{\gamma} \nu_i \mathbf{Z}_i$$

where $\mathbf{Z}_i$ is in $\mathbf{U}(\mathbf{k})$ for $i \leqq \alpha$, and the rest are not. We now use the fact that each of the $\mathbf{Z}_i$, $i > \alpha$, can be written as a sum of switching modes in $\mathbf{U}(\mathbf{k})$ with cost no more than two, and the result follows directly. □

Since the decomposition given by Algorithm Expand is optimal for any decomposition using switching modes in $\mathbf{U}(\mathbf{k})$, its cost can be no worse than the cost of the one given by this procedure. However, we would expect the cost of the decomposition given by Algorithm Expand to not be so bad as this construction suggests.

5. A generalization. It is possible to generalize Algorithm Expand to decompose traffic matrices into sums of other (0, 1)-matrices. Given two vectors of natural numbers $\boldsymbol{\rho} = (\rho_1, \rho_2, \cdots, \rho_m)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_n)$ such that $\sum_{i=1}^m \rho_i = \sum_{j=1}^n \lambda_j$, we consider the problem of decomposing an $m \times n$ traffic matrix into a sum of (0, 1)-matrices which have no more than ρ_1 1's in the first row, ρ_2 1's in the second row, $\cdots$, ρ_m 1's in the mth row, and no more than λ_1 1's in the first column, λ_2 1's in the second column, $\cdots$, λ_n 1's in the nth column. We let $\mathbf{V}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ denote the class of (0, 1)-matrices which satisfy these constraints. If we let $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ denote the class of (0, 1)-matrices which have *exactly* ρ_i 1's in the ith row for all i, and λ_j 1's in the jth column for all j, it is not difficult to generalize Algorithm Expand to decompose $m \times n$ traffic matrices into sums of switching modes in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. There are, however, some restrictions on the traffic matrices which the generalized algorithm will be able to expand. The first restriction comes from the fact that for some $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$, the set $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is empty [4]. If $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is empty, clearly no traffic matrix can be decomposed. The second restriction is that, for certain $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$, $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is nonempty, but there is at least one i and j such that the ijth entry of every switching mode in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is zero. We define such an entry to be a *forced zero entry.* For example, with $\boldsymbol{\rho} = \boldsymbol{\lambda} = (4, 2, 2, 1)$, $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ has several forced zero entries. Clearly, it is impossible to decompose any traffic matrix into a sum of switching modes in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ if it has any nonzero entries which are forced zero entries for $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. We present an algorithm which, given a traffic matrix $\mathbf{T}$, will try to expand it into a traffic matrix $\mathbf{T}'$ such that: $\mathbf{T}'$ dominates $\mathbf{T}$, and $\mathbf{T}'$ can be decomposed into a sum of switching modes in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ which has minimal cost. The algorithm does not assume that either of the above restrictions holds for the particular traffic matrix $\mathbf{T}$ or set $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ in question. However, if either of these restrictions is violated, the violation will be detected by the algorithm (and in the case that the second restriction is violated, at least one nonzero entry in the traffic matrix which is a forced zero entry will be located and reported). Unfortunately, we have been unable to find a bound on the cost of the decomposition generated for these switching modes as compared the optimal decomposition into switching modes in $\mathbf{V}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. Such a bound would correspond to the bound of twice optimal which has been shown for the case of k-permutation switching modes.

5.1. Some preliminaries. We first present the generalization of Theorem 3.1 above for the case of switching modes in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$, whose proof is left to the interested reader.

THEOREM 5.1. *Let* $\mathbf{T}'$ *be a matrix with nonnegative entries. The following conditions are necessary and sufficient for* $\mathbf{T}'$ *to be expressible as a convex sum of* (0, 1)*-matrices in* $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$.

(i) $r_i' = \rho_i x$ *for all* i,

(ii) $c_j' = \lambda_j x$ *for all* j, *and*

(iii) $t_{ij}' \leqq x$ *for all* i *and* j

for some constant x.

The problem here is similar to the one solved in § 3. That is, for a given traffic matrix $\mathbf{T}$, we want to find a matrix $\mathbf{T}'$ such that $\mathbf{T}'$ dominates $\mathbf{T}$; $\mathbf{T}'$ satisfies (i), (ii), and (iii) of Theorem 5.1; and x is as small as possible. In an argument similar to the one given in § 3, we can derive a lower bound on x. From (i) it is seen that x must also be larger than or equal to the ith row sum r_i divided by ρ_i for every i. Similarly, from (ii) it must also be larger than or equal to the jth column sum c_j divided by λ_j for every j. And, from (iii), x must be at least as large as the largest entry in $\mathbf{T}$. We therefore have an initial lower bound on the value x:

$$x = \max_{i,j}\left(t_{ij}, \frac{r_i}{\rho_i}, \frac{c_j}{\lambda_j}\right).$$

In order to perform the expansion, we construct a network which corresponds to the one shown in Fig. 1. There are m edges coming out of the source a, which correspond to the m row sums of $\mathbf{T}$, with the capacity of the ith edge being the total amount that the entries in the ith row must be increased in order to satisfy condition (i) in Theorem 5.1, namely: $\rho_i x - r_i$. Similarly, there are n edges to the sink z, which correspond to the n columns sums of $\mathbf{T}$ with the capacity of the jth edge being the total amount that the entries in the jth column must be increased in order to satisfy condition (ii), namely: $\lambda_j x - c_j$. The mn interior edges correspond to the mn entries of $\mathbf{T}$. The edge from the ith node on the left-hand side to the jth node on the right-hand side corresponds to the ijth entry of $\mathbf{T}$, and its capacity is the largest amount that can be added to this entry without condition (iii) being violated, namely: $x - t_{ij}$. By our choice of x, all of these capacities are nonnegative, so a maximum flow can again be found by known methods.

We define a row saturating flow in this network to be the analogue of the one in § 3, and note that its capacity is $\sum_{i=1}^{m} (\rho_i x - r_i)$. We state without proof the following analogue of Theorem 3.2.

THEOREM 5.2. *There is a* 1-1 *correspondence between row saturating flows in this network, and matrices* $\mathbf{T}'$ *which satisfy the three conditions of Theorem* 5.1.

In the generalized algorithm, as in Algorithm Expand, we start with the initial lower bound for x, construct the network, and try and find a row saturating flow. If one is found, the traffic matrix $\mathbf{T}'$, which is constructed in the same way it was in § 3, is the optimal solution. If one cannot be found, x must be increased, and the procedure iterated.

Again, by permuting rows and columns, any cut in this network can be put in a standard form similar to the one in Fig. 2. The general form for the capacity of such a cut is given by:

$$\sum_{i=t+1}^{m} (\rho_i x - r_i) + \sum_{j=1}^{s} (\lambda_j x - c_j) + \sum_{i=1}^{t} \sum_{j=s+1}^{n} (x - t_{ij}).$$

In a procedure similar to the one in § 3, if a cut F whose capacity is less than that of a row saturating flow is found, x must be increased by some positive value δ so that the capacity of F will be at least as large as a row saturating flow. If we replace x by $x + \delta$ in the cut F, and in the row saturating flow, the following relationship must hold:

$$\sum_{i=t+1}^{m} [\rho_i(x+\delta)-r_i]+\sum_{j=1}^{s}[\lambda_j(x+\delta)-c_j]+\sum_{i=1}^{t}\sum_{j=s+1}^{n}[(x+\delta)-t_{ij}]\geqq\sum_{i=1}^{m}[\rho_i(x+\delta)-r_i]$$

which can be written as

$$\left[\sum_{i=t+1}^{m}\rho_i+\sum_{j=1}^{s}\lambda_j+(n-s)t\right]\delta$$
$$+\left[\sum_{i=t+1}^{m}\rho_i+\sum_{j=1}^{s}\lambda_j+(n-s)t\right]x-\left[\sum_{i=t+1}^{m}r_i+\sum_{j=1}^{s}c_j+\sum_{i=1}^{t}\sum_{j=s+1}^{n}t_{ij}\right]$$
$$\geqq\left[\sum_{i=1}^{m}\rho_i\right]\delta+\sum_{i=1}^{m}(\rho_i x-r_i).$$

If we let

$$A=\sum_{i=1}^{m}\rho_i,$$

$$B=\sum_{i=1}^{m}(\rho_i x-r_i),$$

$$C=\sum_{i=t+1}^{m}\rho_i+\sum_{j=1}^{s}\lambda_j+(n-s)t,$$

$$D=Cx-\left[\sum_{i=t+1}^{m}r_i+\sum_{j=1}^{s}c_j+\sum_{i=1}^{t}\sum_{j=s+1}^{n}t_{ij}\right],$$

the following must hold in order for the expansion to be possible

$$C\delta+D\geqq A\delta+B.$$

By the assumption that the cut is not row saturating, $B > D$, the following three cases can occur.

Case 1: $C > A$. For

$$C\delta+D\geqq A\delta+B$$

or

$$\delta\geqq\frac{B-D}{C-A}$$

to hold, let $\delta = (B - D)/(C - A)$ (which is a positive value), $x = x + \delta$, and construct a new network corresponding to this value for x. We again try to find a flow which will be row saturating for this new value of x. However, by the selection of δ, there is at least one fewer cut which has capacity smaller than a row saturating one.

Case 2: $C = A$. We show that this can occur only if there is at least one nonzero entry in $\mathbf{T}$ which is forced to be zero in every member of $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. To show this and to find at least one such entry, we need the following result.

THEOREM 5.3. *Given a cut of this type where $C = A$, then*

$$\sum_{i=t+1}^{m} \sum_{j=1}^{s} u_{ij} \equiv 0$$

for each **U** *in* $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. (*That is, each entry in this submatrix is a forced zero entry.*)

Proof. Rewrite A and C as

$$A = \sum_{i=1}^{m} \rho_i = \sum_{i=1}^{t} \sum_{j=1}^{s} u_{ij} + \sum_{1=1}^{t} \sum_{j=s+1}^{n} u_{ij} + \sum_{i=t+1}^{m} \sum_{i=1}^{s} u_{ij} + \sum_{i=t+1}^{m} \sum_{s+1}^{n} u_{ij},$$

$$C = \sum_{i=t+1}^{m} \rho_i + \sum_{j=1}^{s} \lambda_j + (n-s)t$$

$$= \sum_{i=t+1}^{m} \sum_{j=1}^{s} u_{ij} + \sum_{i=t+1}^{m} \sum_{j=s+1}^{n} u_{ij} + \sum_{i=1}^{t} \sum_{j=1}^{s} u_{ij} + \sum_{i=t+1}^{m} \sum_{j=1}^{s} u_{ij} + (n-s)t.$$

After equating A and C and simplifying, we get

$$\sum_{i=t+1}^{m} \sum_{j=1}^{s} u_{ij} + (n-s)t = \sum_{i=1}^{t} \sum_{j=s+1}^{n} u_{ij}.$$

But, since

$$(n-s)t \geqq \sum_{i=1}^{t} \sum_{j=s+1}^{n} u_{ij},$$

we must have

$$\sum_{i=t+1}^{m} \sum_{j=1}^{s} u_{ij} \equiv 0. \qquad \square$$

Therefore, we know that every entry in this submatrix must be zero in each switching mode in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. If it can be shown that the additional condition $B > D$ implies that at least one entry in this range is nonzero in the traffic matrix **T**, we know that the expansion is impossible. By hypothesis

$$\sum_{i=1}^{m} (\rho_i x - r_i) > Cx - \left[\sum_{i=t+1}^{m} r_i + \sum_{j=1}^{s} c_j + \sum_{i=1}^{t} \sum_{j=s+1}^{n} t_{ij} \right];$$

however, since $C = A$, this simplifies to

$$\sum_{i=t+1}^{m} r_i + \sum_{j=1}^{s} c_j + \sum_{i=1}^{t} \sum_{j=s+1}^{n} t_{ij} > \sum_{i=1}^{m} r_i.$$

It is easy to show that this can be the case only if

$$\sum_{i=t+1}^{m} \sum_{j=1}^{s} t_{ij} > 0.$$

So we conclude that the expansion is indeed impossible. This submatrix can then be scanned in order to locate one such nonzero entry.

Case 3: $C < A$. We need the following result.

THEOREM 5.4. $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ *is nonempty if and only if* $\sum_{j=1}^{s} \lambda_{y_j} + (n-s)t \geqq \sum_{i=1}^{t} \rho_{x_t}$ *for all subsequences* $\{x_1, x_2, \cdots, x_t\}$ *of* $\{1, 2, \cdots, m\}$, *and* $\{y_1, y_2, \cdots, y_s\}$ *of* $\{1, 2, \cdots, n\}$.

Proof. If $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is nonempty, then for any subset of the rows and columns (which can be assumed to be the first rows and columns by permutation), it is easily shown that this inequality holds, and that in fact equality can hold only if $\sum_{i=t+1}^{m} \sum_{j=1}^{s} z_{ij} = 0$ for some $\mathbf{Z}$ in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$.

To prove the implication in the other direction, assume that $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is empty, and that without loss of generality $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n$, and $\rho_1 \geqq \rho_2 \geqq \cdots \geqq \rho_m$. We state without proof a result from [4] giving a necessary and sufficient condition for the emptiness question for $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. We let ι_k denote a vector of length n which has an initial segment of k 1's followed by $n-k$ 0's. We now consider the matrix $\mathbf{A}$ whose ith rows is ι_{ρ_i}. This matrix is referred to as the *maximal matrix with row sum vector* $\boldsymbol{\rho}$. Let a_j denote the jth column sum of this matrix. It is easily seen that $a_1 \geqq a_2 \geqq \cdots \geqq a_n$. In [4] it is shown that a necessary and sufficient condition for the class $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ to be empty is the existence of some k for which $\sum_{j=1}^{k} \lambda_j > \sum_{j=1}^{k} a_j$, and in particular, we choose the smallest such k. With this choice, it is clear that the $(1, k)$th entry of $\mathbf{A}$ is a one, or there is a contradiction of the selection of k as smallest. We now let t be the largest integer such that the (t, k)th entry of $\mathbf{A}$ is one. With these selections, all of the entries with indices i and j such that $1 \leqq i \leqq t$, and $1 \leqq j \leqq k$ are one.

Now, since $\sum_{j=1}^{k} a_j = \sum_{i=t+1}^{m} \rho_i + kt$, we have $\sum_{j=1}^{k} \lambda_j > \sum_{i=t+1}^{m} \rho_i + kt$, or

$$\sum_{i=1}^{m} \rho_i > \sum_{i=1}^{m} \rho_i - \sum_{j=1}^{k} \lambda_j + \sum_{i=t+1}^{m} \rho_i + kt$$

and from the fact that $\sum_{i=1}^{m} \rho_i = \sum_{j=1}^{n} \lambda_j$, we get

$$\sum_{i=1}^{t} \rho_i > \sum_{j=k+1}^{n} \lambda_j + kt.$$

If we now let $x_i = i$, and $y_j = k + j$ for $1 \leqq j \leqq n-k = s$, we have

$$\sum_{i=1}^{t} \rho_{x_i} > \sum_{j=1}^{s} \lambda_{y_j} + (n-s)t. \qquad \square$$

By assumption, $C < A$, that is,

$$\sum_{i=t+1}^{m} \rho_i + \sum_{j=1}^{s} \lambda_j + (n-s)t < \sum_{i=1}^{m} \rho_i,$$

or

$$\sum_{j=1}^{s} \lambda_j + (n-s)t < \sum_{i=1}^{t} \rho_i,$$

and therefore, $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is empty, and we conclude that the expansion is impossible.

5.2. Generalized Algorithm Expand.

GENERALIZED ALGORITHM EXPAND.

Step 0:

Let $x = \max(\max_{i,j}(t_{ij}), \max_i(r_i/\rho_i), \max_j(c_j/\lambda_j))$.

Step 1:

Construct the network described above, find a maximum flow (minimum cut), and let F denote its capacity.

Step 2:

If $F = \sum_{i=1}^{m} (\rho_i x - r_i)$, we have found a row saturating flow, and we go to Step 5.

Step 3:

Compute the capacities of this minimum cut and of the row saturating flow in the form $C\delta + D$, and $A\delta + B$, respectively, as above.

Step 4:

There are three cases:

$C < A$: $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ is empty by Theorem 5.4. **HALT.**

$C = A$: **T** has nonzero entries which are forced to be zero for every switching mode in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. Find at least one as in Theorem 5.3, report it, and **HALT.**

$C > A$: Let $\delta = (B - D)/(C - A)$, $x = x + \delta$, and return to Step 1.

Step 5:

Increase each of the entries in **T** by the amount of the flow through the corresponding edges in this network, forming the new matrix $\mathbf{T}'$, which at this point satisfies the conditions of Theorem 5.1. **HALT.**

The correctness of the algorithm follows from the above discussion, and termination is insured by the fact that there are only finitely many cuts in the network, and that each iteration of Steps 1–4 satisfies the capacity requirement of at least one more cut. If the algorithm **HALT**ed in Step 5 (that is, a suitable $\mathbf{T}'$ was found), it can then be decomposed using the algorithm presented in [3]. The minimality of the sum of the coefficients ν_i in equation (2.1) is guaranteed by the fact that at each iteration of Step 4, x is incremented by the smallest amount which might allow for the expansion to be possible.

5.3. Analysis of the running time of the generalized algorithm. The analysis of the generalized expansion algorithm is very similar to the analysis of Algorithm Expand, so we shall state without proof that its running time is $O(\mu^5)$, where μ is the larger of m and n, and leave the details to the interested reader.

6. Conclusions. The algorithms presented here provide results related to a problem posed in [2]. That problem corresponds to decomposing a traffic matrix into a convex sum of switching modes in $\mathbf{V}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ which have the further constraint that there is a limit on the total number of ones which appear in each switching mode. It is hoped that the methods presented here will be generalizable to finding optimal solutions to problems of this type. We also hope to be able to bound the performance of the Generalized Expansion Algorithm using switching modes in $\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\lambda})$ in terms of the cost of the optimal decomposition using switching modes in $\mathbf{V}(\boldsymbol{\rho}, \boldsymbol{\lambda})$. This would correspond to the bound for constructing a decomposition using switching modes in $\mathbf{U}(\mathbf{k})$ which has cost no worse than twice that of the optimal decomposition using switching modes in $\mathbf{V}(\mathbf{k})$ which has been obtained here for Algorithm Expand.

REFERENCES

[1] J. Edmonds and R. M. Karp, *Theoretical improvements in algorithmic efficiency for network flow problems*, J. Assoc. Comput. Mach., 19 (1972), pp. 248–264.

[2] I. S. Gopal, G. Bongiovanni, M. A. Bonuccelli, D. T. Tang and C. K. Wong, *An optimal switching algorithm for multibeam satellite systems with variable bandwidth beams*, IBM Research Report RC9006, 1981.

[3] J. L. Lewandowski, C. L. Liu and J. W. S. Liu, *An algorithmic proof of a generalization of the Birkhoff–von Neumann theorem*, J. Algorithms, 7 (1986), pp. 323–330.

[4] H. J. Ryser, *Combinatorial Mathematics*, Mathematical Association of America, Washington, D.C., 1963.

[5] A. K. Sinha, *A model for TDMA burst assignment and scheduling*, COMSAT Tech. Review, vol. 6, no. 7, Fall 1975.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

A DYNAMIC PROGRAMMING APPROACH TO THE DOMINATING SET PROBLEM ON k-TREES*

D. G. CORNEIL† AND J. M. KEIL†‡

Abstract. Dynamic programming has long been established as an important technique for demonstrating the existence of polynomial time algorithms for various discrete optimization problems. In this paper we extend the normal paradigm of dynamic programming to allow a polynomial number of optimal solutions to be computed for each subproblem. This technique yields a polynomial time algorithm for the dominating set problem on k-trees, where k is fixed. It is also shown that the dominating set problem is NP-complete for k-trees where k is arbitrary.

Key words. dominating sets, k-trees, dynamic programming, chordal graphs

AMS(MOS) subject classifications. 05C70, 68R10

1. Introduction.

Dynamic programming is an important technique for the solution of discrete optimization problems. To apply dynamic programming one must represent such a problem by a decision process which proceeds in a series of stages. A dynamic programming algorithm decomposes a problem into a number of smaller subproblems each of which is then further decomposed. Such an algorithm gains its efficiency by avoiding recomputing solutions to common subproblems. For example, a problem with n stages may decompose into several problems with $n - 1$ stages, each of which decomposes into several problems having $n - 2$ stages, etc.

In general, the solution to a subproblem will not be unique. In fact, there are often an exponential number of solutions. For example, a subproblem of the traveling salesman problem is to find a route from the salesman's home city to city i. There are an exponential number of such routes depending upon which intermediate cities are visited. Since it is not clear which of the solutions to this subproblem will grow into the global solution, all solutions must be considered in a dynamic programming algorithm.

If a dynamic programming algorithm is to run in polynomial time it cannot examine an exponential number of subsolutions. Fortunately, there are problems for which keeping a single solution to a subproblem is sufficient. For example, it is sufficient to keep only one of the possibly exponential number of shortest paths from a source vertex to an intermediate vertex when trying to find the shortest path from the source vertex to a destination vertex.

It should be clear that it is desirable to keep as few solutions to a subproblem as possible while ensuring that enough are kept so that at least one is able to form part of the global solution. One way to reduce the number of necessary solutions is to recognize that if two solutions are equivalent in terms of their forming part of a larger solution then only one must be kept. Elmaghraby [3] was the first to recognize this concept of equivalent "states." Note that in the shortest path problem all optimal solutions to a subproblem are equivalent.

In most problems for which polynomial time dynamic programming algorithms have been developed, the solutions to a subproblem form a single equivalence class. The idea that a polynomial time dynamic programming algorithm can be achieved by iden-

* Received by the editors June 28, 1983; accepted for publication (in revised form) December 3, 1986. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4.

‡ Present address, Department of Computational Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0.

tifying a polynomial number of equivalence classes of solutions to subproblems has been recently exploited on some polygonal decomposition problems in the field of computational geometry [8]. In this paper we apply this idea of keeping a polynomial number of solutions to each subproblem of a dynamic programming formulation to a problem in graph theory.

Let us now turn to the problem in question. D, a set of vertices in a graph $G(V, E)$, is a *dominating set* if for every vertex in $V \backslash D$ there exists an adjacent vertex in D. As reported in [5] the h-dominating set problem, i.e., determining if a graph has a dominating set of cardinality $\leqq h$, is NP-complete for arbitrary graphs. Recently the problem has been studied for various classes of *chordal graphs* (namely, graphs where every cycle of length greater than three has a chord). In particular, Booth and Johnson [1] have studied domination on the following hierarchy of chordal graphs: chordal graphs $\supsetneq$ undirected path graphs $\supsetneq$ directed path graphs $\supsetneq$ interval graphs. They produced a linear time algorithm for the h-domination set problem on directed path graphs and also established the NP-completeness of the problem on undirected path graphs. Previous work on other classes of chordal graphs includes polynomial time algorithms for trees [2] and strongly chordal graphs [4].

A *k-tree* is defined recursively as follows: K_k is a k-tree, and if G is a k-tree then so is G' the graph formed by adding a new vertex to G and making it adjacent to all vertices in a K_k in G. An example of a 2-tree is presented in Fig. 1. Note that a 1-tree is a tree in the normal sense. Furthermore, k-trees are chordal and are specific examples of hookup graphs defined in [10].

Another class of graphs which is related to k-trees, but is not chordal, is the class of series-parallel graphs. The class of series-parallel graphs includes the class of 2-trees but has no strong relationship with other k-trees. Recently Kikuno et al. [9] have developed a linear time algorithm for the h-domination set problem on series-parallel graphs.

In this paper we will examine domination problems on another hierarchy of chordal graphs, namely, chordal graphs $\supsetneq$ k-trees with unbounded k $\supsetneq$ k-trees with bounded k $\supsetneq$ trees. We will use our dynamic programming technique to outline a polynomial time algorithm for k-trees with bounded k. We will also demonstrate the NP-completeness of the h-dominating set problem on k-trees with unbounded k.

2. *k*-trees; *k* bounded. In this section we will show how the tree-like structure of k-trees can be exploited in a dynamic programming algorithm for the domination problem. A *simplicial* vertex is one for which the set of neighbours forms a clique. A graph has a *perfect elimination scheme* if there exists an order of eliminating the vertices such that each vertex is simplicial at the time it is eliminated. A graph is chordal iff it has a perfect

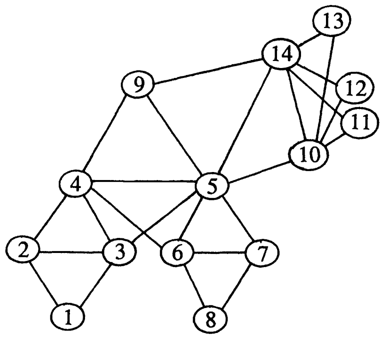


FIG. 1

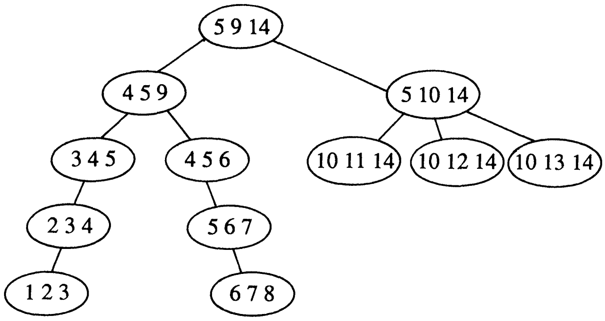


FIG. 2

elimination scheme [7]. This perfect elimination scheme indicates that the maximum cliques of a chordal graph interlock in a very tree-like way. This clique tree [6] can be constructed in linear time [11]. See Fig. 2 for an example of a clique tree.

One obvious approach to the dominating set problem on k-trees is to attempt to use the clique tree to generalize the tree algorithm of Cockayne et al. [2]. The tree algorithm starts at the leaves and delays putting a vertex in the dominating set until such a move is forced. For k-trees we would first find minimum dominating sets of the subgraphs corresponding to the leaves in the clique tree and then form minimum dominating sets for the subgraphs corresponding to larger and larger subtrees of the clique tree. This does not work since a minimum dominating set on a sub-k-tree does not necessarily belong to any minimum dominating set of the k-tree. The existence of families of k-trees where the number of dominating sets grows exponentially with n rules out an exhaustive search strategy to achieve a polynomial time algorithm. Instead we use the previously mentioned idea of keeping a polynomial number of dominating sets of each sub-k-tree in a dynamic programming formulation.

2.1. Terminology. Given a k-tree G, an associated clique tree T and a $(k+1)$-clique C of G we define $G(C)$ to be the subgraph of G corresponding to the cliques in the subtree of T rooted at C. In the example in Fig. 2, if $C = \{5,10,14\}$ then $G(C)$ corresponds to the graph induced on the nodes $\{5,10,11,12,13,14\}$. Furthermore, we let $C_1, C_2, \cdots, C_l$ denote the children of C in the tree T. In our example the children of C are $\{10,11,14\}$, $\{10,12,14\}$ and $\{10,13,14\}$. If C_i is a child of C, then x_i denotes the vertex $\in C\backslash C_i$ and $\bar{x}_i$ denotes the vertex $\in C_i\backslash C$. In Fig. 2, if $C_i = \{10,12,14\}$, then $x_i = 5$ and $\bar{x}_i = 12$.

The algorithm will calculate a minimum dominating set of G by performing a dynamic programming algorithm on T and for each clique C of T calculating the following two sets of quantities. For every subset S of C we determine the following:

(i) $DS(G(C), S)$—a minimum sized dominating set of $G(C)$ such that the dominating set contains S.

(ii) $\tilde{DS}(G(C), S)$—a minimum sized dominating set of $G(C)$ assuming that the vertices in S are already dominated.

Clearly if R is the root of T then $G(R) = G$ and $DS(G(R), \emptyset)$ is a minimum sized dominating set of G.

2.2. The algorithm. Before describing the algorithm we present a procedure "GENERATE" used in the algorithm. Given a clique C with children $\{C_i\}$ $1 \leqq i \leqq l$ and a subset S of C, GENERATE attempts to compute V (a union of $\tilde{DS}$ sets) which dominates $G(C)\backslash S$. A successful computation of V is determined by a mapping of the vertices in

$C \backslash S$ to the cliques $C_1, \cdots, C_l$. If V_i denotes the set of vertices of $C \backslash S$ assigned to C_i, and if $|\tilde{DS}(G(C_i), (C \cap C_i) \backslash V_i)| = |\tilde{DS}(G(C_i), C \cap C_i)|$, then $V = \bigcup_{i=1}^{l} \tilde{DS}(G(C_i), (C \cap C_i) \backslash V_i)$. If such a V does not exist, then GENERATE fails. The following description of GENERATE allows each V_i to be chosen nondeterministically. As discussed in § 2.5, a deterministic implementation could result in an exponential time bound.

```
Procedure GENERATE ($C$, $\{C_i\}$, $S$)
    input    $C$ with children $\{C_i\}$, $1 \leqq i \leqq l$
             $S \subseteq C$
    output   $V$, the union of $l$ $\tilde{DS}$ sets or "Failure"
                 $V = \bigcup_{i=1}^{l} \tilde{DS}(G(C_i), (C \cap C_i) \backslash V_i)$
such that (i)    $|\tilde{DS}(G(C_i), (C \cap C_i) \backslash V_i)| = |\tilde{DS}(G(C_i), C \cap C_i)|$,
          (ii)   $V_i \subseteq C \cap C_i$,
          (iii)  $V$ dominates $C \backslash S$.
If no such set exists return "failure".
```

We now describe the algorithm. The computation progresses from the leaves of T to the root R. In all cases the DS sets are calculated before the $\tilde{DS}$ sets. Throughout the algorithm MIN refers to choosing the set with the smallest cardinality.

Step I. C is a leaf of T
- (a) Compute $DS(G(C), S)$, $\forall S \in 2^C$
  - (1) If $S \neq \emptyset$, $DS(G(C), S) = S$
  - (2) If $S = \emptyset$, $DS(G(C), \emptyset) = \{x\}$ where x is an arbitrary element of C.
- (b) Compute $\tilde{DS}(G(C), S)$, $\forall S \in 2^C$
  - (1) If $S = C$, $\tilde{DS}(G(C), C) = \emptyset$
  - (2) If $S \neq C$, $S \neq \emptyset$, $\tilde{DS}(G(C), S) = \{x\}$ where x is an arbitrary element of C.
  - (3) If $S = \emptyset$, $\tilde{DS}(G(C), \emptyset) = DS(G(C), \emptyset)$.

Step II. C is not a leaf of T. (We assume that the DS and the $\tilde{DS}$ sets have been computed for all C_i, $1 \leqq i \leqq l$.)
- (a) Compute $DS(G(C), S)$, $\forall S \in 2^C$. This progresses in nonincreasing order of $|S|$ so that in calculating $DS(G(C), S)$ we may assume that $DS(G(C), S')$ is known $\forall S'$, $|S'| > |S|$.
  - (1) If $S \neq \emptyset$,

$$DS(G(C), S) = \text{MIN} \begin{cases} S \bigcup_{i=1}^{l} \begin{cases} \{DS(G(C_i), S \cap C_i)\} & \text{if } S \cap C_i \neq \emptyset, \quad \text{(a)} \\ \{\tilde{DS}(G(C_i), C \cap C_i)\} & \text{if } S \cap C_i = \emptyset, \quad \text{(b)} \end{cases} \\ \{DS(G(C), S \cup \{x\})\} \quad \forall x \in C \backslash S. & \text{(c)} \end{cases}$$

  - (2) If $S = \emptyset$,

$$DS(G(C), \emptyset) = \text{MIN} \begin{cases} \text{GENERATE}\,(C, \{C_i\}, \emptyset), \\ \{DS(G(C), \{x\})\} \quad \forall x \in C. \end{cases}$$

- (b) Compute $\tilde{DS}(G(C), S)$ $\forall S \in 2^C$. (We assume that $DS(G(C), \emptyset)$ has already been computed.)

(1) If $S \neq \emptyset$,

$$\tilde{DS}(G(C), S) = \text{MIN}\begin{cases} \text{GENERATE}(C, \{C_i\}, S), \\ DS(G(C), \emptyset). \end{cases}$$

(2) If $S = \emptyset$, $\tilde{DS}(G(C), \emptyset) = DS(G(C), \emptyset)$.

2.3. Example. Let us consider how the algorithm would progress on the 2-tree shown in Fig. 1. At a particular point the sets DS and $\tilde{DS}$ are available (see Table 1). Using the rules in the algorithm the sets DS and $\tilde{DS}$ are computed as shown in Table 2. After the sets DS and $\tilde{DS}$ are available for $G(\{5,10,14\})$ we compute the DS and $\tilde{DS}$ for $G(\{5,9,14\})$ (see Table 3).

2.4. Proof of correctness. We now show that the set produced by $DS(G(R), \emptyset)$ is a minimum sized dominating set of G. First we state some lemmas which will be used in the proof.

LEMMA 1. *Let C be a clique with children $C_1, C_2, \cdots, C_l$. For any i, j, $1 \leqq i, j \leqq l$, $i \neq j$. $G(C_i) \cap G(C_j) \subsetneqq C$.*

Proof. In fact $G(C_i) \cap G(C_j) = C \backslash (\{x_i\} \cup \{x_j\})$. This follows from the definitions of k-trees and clique trees. □

LEMMA 2. *Given $x \in G(C_i) \backslash C$, the vertices of G adjacent to x must lie in $G(C_i)$.*

Proof. This follows from the definition of k-trees. See Fig. 3. □

LEMMA 3. *Let X be a dominating set of $G(C)$.*

(i) *If $X \cap (C \cap C_i) \neq \emptyset$, then $X \cap G(C_i)$ dominates $G(C_i)$.*

(ii) *If $X \cap C = \{x_i\}$, then $(X \cap G(C_i)) \cup \{x_i\}$ dominates $G(C_i)$.*

(iii) *If $X \cap C = \emptyset$, then $X \cap G(C_i)$ dominates $G(C_i) \backslash (C \cap C_i)$.*

Proof. From Lemma 2, any vertex in $G(C_i) \backslash C$ can only be dominated by vertices in $G(C_i)$ and thus the vertices in $X \cap G(C_i)$ dominate the vertices in $G(C_i) \backslash (C \cap C_i)$. In case (i) the only other vertices are those in $C_i \cap C$; however, they are dominated by any vertex in $X \cap (C \cap C_i)$. In case (ii) x_i dominates the vertices in $C_i \cap C$. □

LEMMA 4. *If GENERATE $(C, \{C_i\}, S)$ does not return "failure," then the set produced by GENERATE $(C, \{C_i\}, S)$ dominates $G(C) \backslash S$.*

Proof. From condition (iii) of GENERATE, V the output of the procedure dominates $C \backslash S$. Since each of the $\tilde{DS}(G(C_i), (C \cap C_i) \backslash T_i)$ dominates $G(C_i) \backslash C$ the lemma is proved. □

THEOREM 1. *The $DS(G(R), \emptyset)$ calculated by the algorithm is a minimum dominating set of G.*

TABLE 1

	$C = \{3,4,5\}$			$C = \{4,5,6\}$	
S	$DS(G(C), S)$	$\tilde{DS}(G(C), S)$	S	$DS(G(C), S)$	$\tilde{DS}(G(C), S)$
$\emptyset$	$\{3\}$	$\{3\}$	$\emptyset$	$\{6\}$	$\{6\}$
$\{3\}$	$\{3\}$	$\{3\}$	$\{4\}$	$\{4,8\}$	$\{6\}$
$\{4\}$	$\{4,1\}$	$\{3\}$	$\{5\}$	$\{5,8\}$	$\{6\}$
$\{5\}$	$\{5,1\}$	$\{2\}$	$\{6\}$	$\{6\}$	$\{6\}$
$\{3,4\}$	$\{3,4\}$	$\{3\}$	$\{4,5\}$	$\{4,5,8\}$	$\{8\}$
$\{3,5\}$	$\{3,5\}$	$\{2\}$	$\{4,6\}$	$\{4,6\}$	$\{6\}$
$\{4,5\}$	$\{4,5,1\}$	$\{2\}$	$\{5,6\}$	$\{5,6\}$	$\{6\}$
$\{3,4,5\}$	$\{3,4,5\}$	$\{2\}$	$\{4,5,6\}$	$\{4,5,6\}$	$\{8\}$

TABLE 2

S	$C = \{4,5,9\}$	
	$DS(G(C), S)$	$\tilde{D}S(G(C), S)$
$\varnothing$	{4,1,8}	{4,1,8}
{4}	{4,1,8}	{4,1,8}
{5}	{5,1,8}	{4,1,8}
{9}	{9,2,8}	{2,6}
{4,5}	{4,5,1,8}	{4,1,8}
{4,9}	{4,9,1,8}	{3,8}
{5,9}	{5,9,1,8}	{2,8}
{4,5,9}	{4,5,9,1,8}	{3,6}

Proof. To prove this it is sufficient to show that the DS and $\tilde{D}S$ sets determined by the algorithm do in fact satisfy their definitions. This is done by induction and follows the order of calculation of the algorithm.

If C is a leaf, then it is clear that the algorithm is correct. We now assume that C is not a leaf and that the DS and $\tilde{D}S$ sets have been calculated correctly for all C_i, the children of C.

Part 1. $DS(G(C), S)$. Throughout the proof we let D denote the DS set computed by the algorithm. The proof proceeds by induction on $|S|$. For the base case $S = C$ and the only applicable clause is (a). In other words, we wish to show that $D = C \cup_{i=1}^{l} \{DS(G(C_i), C \cap C_i)\}$ is a minimum dominating set of $G(C)$ which contains C. It is obvious that D is a dominating set of $G(C)$ and that $C \subseteq D$; we now establish the minimum cardinality of D. Assume this is not true and there exists X s.t. $C \subseteq X \subseteq G(C)$; X is a dominating set of $G(C)$ and $|X| < |D|$. Since $X \cap (C \cap C_i) \neq \varnothing$, Lemma 3 states that $X \cap G(C_i)$ dominates $G(C_i)$, and thus $X \cap G(C_i)$ is a possible candidate for $DS(G(C_i), C \cap C_i)$, for all i. To show that $|X| \geqq |D|$ we first observe from Lemma 1 that $(X \cap G(C_i)) \cap (X \cap G(C_j)) \subsetneq C$, $1 \leqq i, j \leqq l$, $i \neq j$. Thus vertices in $X \backslash C$ must belong to some $G(C_i) \backslash C$ and, from Lemma 2, cannot dominate any vertex in $G(C_j) \backslash C$, $j \neq i$. In other words $[(X \cap G(C_i)) \backslash C] \cap [(X \cap G(C_j)) \backslash C] = \varnothing$, $1 \leqq i, j \leqq l$, $i \neq j$, thereby showing the independence of the $(X \cap G(C_i)) \backslash C$, $i = 1, \cdots, l$. Thus

$$|X| = |C| + \sum_{i=1}^{l} |(X \cap G(C_i)) \backslash C|.$$

From the minimality assumption of $|DS(G(C_i), C \cap C_i)|$ and the fact that $X \cap G(C_i)$ is

TABLE 3

S	$C = \{5,10,14\}$		S	$C = \{5,9,14\}$	
	$DS(G(C), S)$	$\tilde{D}S(G(C), S)$		$DS(G(C), S)$	$\tilde{D}S(G(C), S)$
$\varnothing$	{10}	{10}	$\varnothing$	{14,2,8}	{14,2,8}
{5}	{5,10}	{10}	{5}	{5,10,1,8}	{14,2,8}
{10}	{10}	{10}	{9}	{9,10,2,8}	{2,6,10}
{14}	{14}	{10}	{14}	{14,2,8}	{14,2,8}
{5,10}	{5,10}	{10}	{5,9}	{5,9,10,1,8}	{2,8,10}
{5,14}	{5,14}	{10}	{5,14}	{5,14,1,8}	{14,2,8}
{10,14}	{10,14}	{10}	{9,14}	{9,14,2,8}	{2,6,10}
{5,10,14}	{5,10,14}	{10}	{5,9,14}	{5,9,14,1,8}	{2,8,10}

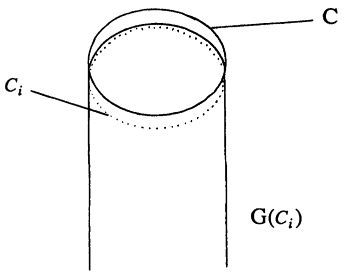


FIG. 3

a possible candidate for $DS(G(C_i), C \cap C_i)$, we conclude that $|X| \geqq |D|$ contradicting $|X| < |D|$.

We now assume that $DS(G(C), S)$ is a minimum dominating set of $G(S)$ containing S, for all $S \in 2^C$ such that $|S| > h$ and show that the algorithm computes a legitimate $DS(G(C), S)$ for $|S| = h$. This proof involves 2 cases.

Case 1. $S \neq C$, $|S| \geqq 1$. It is clear that regardless of the choice taken by the MIN operator, the set D contains S and is a dominating set. To show it is of minimum cardinality we assume to the contrary that there exists X such that X is a dominating set of $G(C)$, $S \subseteq X$ and $|X| < |D|$.

If there exist $z \in X \cap (C \backslash S)$, then X is a dominating set of $G(C)$ containing $S \cup \{z\}$ contradicting the fact that $|D| \leqq \text{MIN}_{X \in C \backslash S} \{|DS(G(C), S \cup \{x\})|\}$ and the inductive assumption that all of these dominating sets are of minimum cardinality.

Thus we may assume that $X \cap (C \backslash S) = \emptyset$; in other words, $X \cap C = S$. If $|S| > 1$ or $S \cap C_i \neq \emptyset$ then by Lemma 3(i) $X \cap G(C_i)$ is a dominating set of $G(C_i)$. In these cases $X \cap G(C_i)$ is a possible candidate for $DS(G(C_i), S \cap C_i)$. If $|S| = 1$ and $S \cap C_i = \emptyset$ (i.e., $S = \{x_j\}$ and clause (b) is chosen), then by Lemma 3(ii) $X \cap G(C_i)$ is a possible candidate for $\tilde{D}S(G(C_i), C \cap C_i)$. A similar argument to that for the base case shows that $[(X \cap G(C_i)) \backslash S] \cap [(X \cap G(C_j)) \backslash S] = \emptyset$, $1 \leqq i, j \leqq l$, $i \neq j$, thereby establishing the independence of the $(X \cap G(C_i)) \backslash S$, $i = 1, \cdots, l$. This shows that

$$|X| = |S| + \sum_{i=1}^{l} |(X \cap G(C_i)) \backslash S|.$$

As in the base case, the minimality assumptions of the $DS(G(C_i), S \cap C_i)$ and the $\tilde{D}S(G(C_i), C \cap C_i)$ imply that $|X| \geqq |D|$ contradicting $|X| < |D|$.

Case 2. $S = \emptyset$. Using Lemma 4, it is clear that regardless of the clause chosen, D is a dominating set of $G(C)$. To show that it is of minimum cardinality we again assume that there exists a smaller such set X. If $z \in X \cap C$ then X is a dominating set of $G(C)$ containing $\{z\}$, contradicting the fact that $|D| \leqq \text{MIN}_{X \in C} \{|DS(G(C), \{x\})|\}$ and the inductive assumption that all of these dominating sets are of minimum cardinality.

Thus we may assume that $X \cap C = \emptyset$. From Lemma 3(iii) we know that $X \cap G(C_i)$ dominates $G(C_i) \backslash (C \cap C_i)$. Since $X \cap C = \emptyset$, $[X \cap G(C_i)] \cap [X \cap G(C_j)] = \emptyset$, $i \neq j$ and $|X| = \sum_{i=1}^{l} |X \cap G(C_i)|$. By the inductive assumption, $|X \cap G(C_i)| \geqq |\tilde{D}S(G(C_i), C \cap C_i)|$. From (i) in procedure GENERATE, $|\tilde{D}S(G(C_i), C \cap C_i)| = |\tilde{D}S(G(C_i), (C \cap C_i) \backslash V_i)|$ for all $V_i \subseteq C \cap C_i$ and thus

$$|V| \leqq \sum_{i=1}^{l} |\tilde{D}S(G(C_i), (C \cap C_i) \backslash V_i)| = \sum_{i=1}^{l} |\tilde{D}S(G(C_i), C \cap C_i)|$$

thereby showing that $|X| \geqq |V| = |D|$, contradicting $|X| < |D|$.

Part 2. $\widetilde{DS}(G(C, S)$. We now show that the $\widetilde{DS}$ sets are determined correctly. As in Part 1 we assume that the $\widetilde{DS}$ sets have been calculated correctly for all C_i, the children of C. We may also assume that $DS(G(C), \varnothing)$ has been accurately determined. We let $\tilde{D}$ denote the set produced by the algorithm. From Lemma 4 it is clear that regardless of the clause used, $\tilde{D}$ is a dominating set of $G(C)\setminus S$. We now show it is of minimum cardinality by assuming the existence of a smaller such set $\tilde{X}$. If $\tilde{X} \cap C \neq \varnothing$ then $\tilde{X}$ not only is a candidate for $\widetilde{DS}(G(C), S)$ but is also a candidate for $DS(G(C), \varnothing)$ since $\tilde{X}$ dominates S. Since $DS(G(C), \varnothing)$ is of minimum cardinality by assumption, we must assume that $\tilde{X} \cap C = \varnothing$. The rest of the proof follows exactly the proof at the end of Part 1, Case 2. □

2.5. Analysis. Let us first consider the time complexity of a call to a deterministic implementation of procedure GENERATE. There are $l^{|C\setminus S|}$ different assignments of the vertices in $C\setminus S$ to the l children of C. For each of the at most l^{k+1} assignments $O(l2^{k+1})$ work is performed to check for success. Altogether a call to GENERATE is performed in $O(l^{k+2}2^{k+1})$ time in the worst case.

At each node of the clique tree $O(2^{k+1})$ $DS(G(C), S)$ and $\widetilde{DS}(G(C), S)$ sets are computed. Each of these computations will require a call to procedure GENERATE or l searches (each requiring time $O(2^{k+1})$) for a $DS(G(C_i), X)$ or $\widetilde{DS}(G(C_i), X)$. Therefore $O(l^{k+2}2^{2k+2})$ time is spent at each node of the clique tree. Since there are $O(n)$ maximum cliques in a k-tree and thus $O(n)$ nodes in the clique tree of a k-tree, the algorithm finds all required DS and $\widetilde{DS}$ sets in $O(n^{k+3}2^{2k+2})$ time. Thus if k is fixed, we have a polynomial time algorithm for the domination problem on k-trees. It should be noted that the exponent of n, namely, $k + 3$, can be reduced to $k + 2$ by means of a careful implementation.

Furthermore, it is important to note that the same algorithm may be modified to give a polynomial time algorithm for the dominating set problem on chordal graphs where the clique size is bounded by a constant.

3. *k*-trees; *k* unbounded.

In this section we establish the NP-completeness of the h-dominating set problem on k-trees for arbitrary k. h-domination on general graphs was proved NP-complete using a reduction from the h-vertex cover problem [5]. We also use this problem to establish NP-completeness on k-trees with unbounded k. In particular, we show that the h-vertex cover problem for an arbitrary graph G may be polynomially reduced to the h-dominating set problem for G' where G' is an n-tree constructed from G and $n = |G|$. Thus we have the following theorem.

THEOREM 2. *The h-dominating set problem on k-trees with arbitrary k is* NP-*complete.*

Proof. We construct G' from G as follows. Each vertex V_i of G is represented by a vertex V'_i in a central complete subgraph C of G'. Each edge e_{ij} of G is represented by $n - 1$ vertices $V'_{ij1}, V'_{ij2}, \cdots, V'_{ijn-1}$. Vertex V'_{ij1} is made adjacent to each vertex of C. Vertex V'_{ij2} is made adjacent to V'_{ij1} and a subset S_2 of $n - 1$ vertices of C that includes V'_i and V'_j. Vertex V'_{ijl} is made adjacent to $V'_{ij1}, V'_{ij2}, \cdots, V'_{ijl-1}$ and a subset S_l of S_{l-1} of size $n - l + 1$ that includes the vertices V'_i and V'_j. Finally, V'_{ijn-1} is made adjacent to $V'_{ij1}, V'_{ij2}, \cdots, V'_{ijn-2}$ and V'_i and V'_j. Note that $|V'| = n + (n - 1)|E|$. It is clear that G' can be constructed from G in polynomial time. Figure 4 shows an example of the construction of G' from G. We claim that G has an h-vertex cover if and only if G' has an h-dominating set.

Each vertex cover for G corresponds to a dominating set for G'. Since the central part of G' is complete, all vertices of G' that represent vertices of G are dominated by vertices in G' that correspond to the vertex cover in G. Every edge of G is incident to a vertex in the cover for G; therefore by our construction each vertex in G' that represents

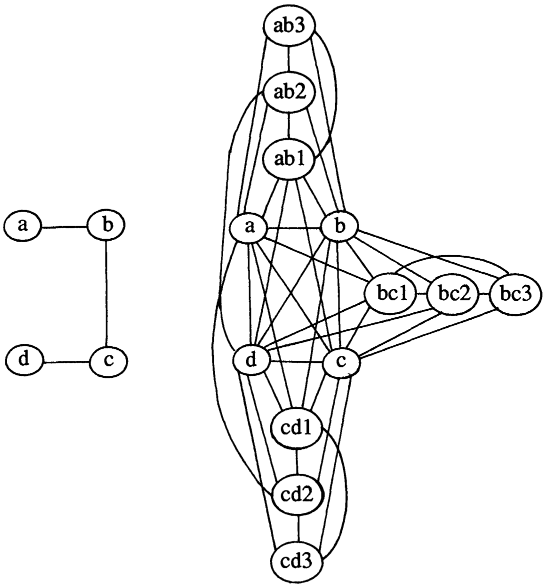


FIG. 4. *The construction of G' from G.*

an edge in G is dominated by a vertex in G' that corresponds to a vertex in the vertex cover for G.

Given a dominating set for G', we can assume without loss of generality that it contains only vertices in the central complete subgraph. If the dominating set contained a vertex V_{ijl}, then V_{ijl} can be replaced by either V_i or V_j and the set will still be dominating. Since all the vertices in G' that correspond to edges in G are adjacent to a vertex in the dominating set, the vertices in G corresponding to the vertices in the dominating set for G' will form a vertex cover for G. That G' is an n-tree is evident from the construction. □

REFERENCES

[1] K. S. BOOTH AND J. H. JOHNSON, *Dominating sets in chordal graphs*, SIAM J. Comput., 11 (1982), pp. 191–199.

[2] E. COCKAYNE, S. GOODMAN AND S. HEDETNIEMI, *A linear algorithm for the domination number of a tree*, Inform. Process. Lett., 4 (1975), pp. 41–44.

[3] S. E. ELMAGHRABY, *The concept of "State" in discrete dynamic programming*, J. Math. Anal. Appl., 29 (1970), pp. 523–557.

[4] M. FARBER, *Application of linear programming duality to problems involving independence and domination*, Ph.D. thesis, Rutgers Univ., New Brunswick, NJ, available as TR81-13, Simon Fraser University, Burnaby, British Columbia, Canada.

[5] 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.

[6] F. GAVRIL, *Algorithms for minimum coloring, maximum clique, minimum covering by cliques, and maximum independent set of a chordal graph*, SIAM J. Comput., 1 (1972), pp. 180–187.

[7] M. C. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[8] J. M. KEIL, *Decomposing polygons into simpler components*, SIAM J. Comput., 14 (1985), pp. 799–817.

[9] T. KIKUNO, N. YOSHIDA AND Y. KAKUDA, *A linear algorithm for the domination number of a series parallel graph*, Discrete Appl. Math., 5 (1983), pp. 299–311.

[10] M. M. KLAWE, D. G. CORNEIL AND A. PROSKUROWSKI, *Isomorphism testing in hookup classes*, this Journal, 3 (1982), pp. 260–274.

[11] D. J. ROSE, R. E. TARJAN AND G. S. LUEKER, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

003

THE NULL SPACE PROBLEM II. ALGORITHMS*

THOMAS F. COLEMAN† AND ALEX POTHEN‡

Abstract. The null space problem is that of finding a sparsest basis for the null space (*null basis*) of an underdetermined matrix. This problem was shown to be NP-hard in Coleman and Pothen (this Journal, 7 (1986), pp. 527–537). In this paper we develop heuristic algorithms to find sparse null bases. A basis is computed by columns, i.e., by finding a null vector linearly independent of those previously obtained. The algorithms to compute null vectors have two phases. In the first combinatorial phase, a minimal dependent set of columns is identified by finding a matching in the bipartite graph of the matrix. In the second numerical phase, nonzero coefficients in the null vector are computed from this dependent set.

We have designed two algorithms: the first computes a fundamental basis (one with an embedded identity matrix), and the other, a triangular basis (one with an upper triangular matrix). We describe implementations of our algorithms and provide computational results on several large sparse constraint matrices from linear programs. Both algorithms find null bases which are quite sparse, have low running times, and require small intermediate storage. The triangular algorithm finds sparser bases at the expense of greater running times. We believe that this algorithm is an attractive candidate for large sparse null basis computations.

Key words. null basis, null space, sparse matrix, bipartite graph, matching, linear programming, nonlinear programming

AMS(MOS) subject classifications. 05, 15, 49, 65, 68

1. Introduction.

Currently successive quadratic programming is the most popular method to solve constrained nonlinear optimization problems. The quadratic programming subproblems are often solved by numerically stable null space algorithms. Thus designing efficient null space algorithms for large scale optimization problems is an area of intense research effort at present. One concern is that these algorithms require a sparse representation of the null space of the constraint matrix.

Let A be a $t \times n$ matrix of rank t. The Null Space Problem (NSP) (Pothen (1984), Coleman and Pothen (1986a)) is to find a basis N, with the fewest nonzeros, for the null space of A. For brevity, a basis for the null space will be called a *null basis*, and a column of a null basis will be called a *null vector*.

Two representations for the null basis N have been used so far in optimization algorithms. Wolfe (1962) proposed permuting the columns of A to obtain a $t \times t$ nonsingular matrix M such that $A = (MU)$, so that

$$N = \begin{pmatrix} B \\ I_{n-t} \end{pmatrix} \tag{1.1}$$

where $B \equiv -M^{-1}U$, and I_{n-t} is the $(n-t)$-dimensional identity matrix. We will call such a basis a *fundamental* null basis. An *explicit representation* of N is one in which the nonzeros in N are stored. In practice, N is represented *implicitly* by storing the LU factors of M, and a matrix-vector product such as Np is computed by solving a system of equations involving M. In the second representation, an LQ factorization of A is computed, and

* Received by the editors May 13, 1986; accepted for publication (in revised form) December 11, 1986. This research was supported in part by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under contract DE-AC02-83-ER10369. A preliminary version of this work was presented at the SIAM Conference on Applied Linear Algebra at Rayleigh, North Carolina, in April 1985.

† Department of Computer Science, Cornell University, Ithaca, New York 14853.

‡ Department of Computer Science, The Pennsylvania State University, University Park, Pennsylvania 16802.

the last $n - t$ columns of Q form an orthogonal null basis for A. This scheme is impractical for large scale problems since Q is likely to be quite dense.

A context in which an explicit representation of the null basis is required concerns the optimization of a nonlinear function subject to linear constraints. A null space method demands the matrix N^THN, where H is the current Hessian or an approximation to it. If a subroutine to compute the gradient of the objective function is available, then it is possible to obtain a good approximation to HN by $n - t$ extra gradient evaluations (Gill, Murray and Wright (1981)), provided N is explicit. If N is explicit, and HN is sparse, then, in general, many fewer gradient evaluations will be needed if a sparse finite difference scheme is used (Coleman and Moré (1983), (1984)).

A second context arises from the recent work of Goldfarb and Mehrotra (1985) and Shanno and Marsten (1985) which extends Karmarkar's algorithm for linear programming. The crucial computational step in these works is the solution of large sparse linear least squares problems of the form

$$DNw = b,$$

where D is a diagonal matrix, and N is a null basis of the constraint matrix. Both groups suggest solving the linear systems using pre-conditioned conjugate gradients; however, a host of pre-conditioning strategies is lost if N is not explicit. For example, diagonal, incomplete Cholesky, and chordal (Coleman (1986)) pre-conditioners, all require an explicit null basis N. (Thapa (1984) discusses a variety of pre-conditioners available for optimization problems: most require explicit matrices.)

Hence, for the rest of this paper, we restrict ourselves to the study of sparse explicit representations of null bases.

Previous work by others. Recently much work has been done on computing sparse null bases. The "turnback" method for computing a null basis with a profile structure for equilibrium matrices in structural analysis was proposed by Topcu (1979). Kaneko, Lawo and Thierauf (1982) interpreted this algorithm from a matrix factorization point of view. Berry, Heath, Kaneko, Lawo, Plemmons and Ward (1985) refined this algorithm, implemented it using profile data structures, and tested it on several structural problems. Berry and Plemmons (1985) have implemented this algorithm on a HEP multiprocessor.

The turnback algorithm computes a QR factorization of A to identify a set of $(n - t)$ start columns. These are columns which are identified as linearly dependent in the factorization. Hence there is a null vector containing a start column and columns numbered lower than it in the matrix. Each null vector is computed by an algorithm which maintains a set of active columns, initially containing only a start column. Lower numbered columns are added to the active set, one by one, and a QR factorization of the active set is maintained. When the active set becomes dependent, the columns correspond to the nonzero components of a null vector. If the dependence involves the start column, the null vector is accepted. If not, the dependent column is rejected from the active set, and the process continued.

Null bases obtained by turnback are not fundamental; they have an embedded upper triangular matrix U_{n-t} of dimension $(n - t)$ with nonzero diagonal elements. Thus

$$N = \begin{pmatrix} B \\ U_{n-t} \end{pmatrix}, \tag{1.2}$$

and we call such bases *triangular* null bases.

Gilbert and Heath (1987) have implemented several algorithms to compute sparse null bases. One of these is the turnback algorithm using general sparse data structures from Sparspak (George and Liu (1981)). Another is a matching based algorithm that computes triangular bases; we discuss this algorithm in § 4 of this paper.

Previous work by the authors. NSP was formulated in Coleman and Pothen (1986a). We briefly summarize definitions and major results from that paper that will be useful here.

A null vector of A can be obtained from a linearly dependent set of columns. We call such a set a *dependent set*. The coefficients of the linear combination correspond to the values of the nonzero components of the null vector. A minimal dependent set of columns of A is a *circuit*. We proved that only null vectors that correspond to circuits could be columns in a sparsest null basis.

Sparsest null bases were characterized by a greedy algorithm that augmented a partial basis by a sparsest null vector independent of those previously chosen. Despite this result, finding a sparsest null basis is computationally an intractable problem since it is NP-hard. Computing a sparsest fundamental null basis is also NP-hard.

We addressed the question if sparsest null bases could be characterized to have some particular zero-nonzero structure (*structure*). It is known that a sparsest null basis may not be fundamental. We showed that a set of k vectors is linearly independent for all possible numeric values of its nonzeros if and only if it has an embedded upper triangular matrix of dimension k. Yet we do not know if we can always restrict a sparsest basis to be triangular. Nevertheless, restriction of the structures to fundamental and triangular bases makes it easy to ensure linear independence of the null vectors.

The relation between a triangular and a fundamental basis is an interesting one. Since a triangular null basis has the structure in (1.2), if A is partitioned to conform to N as $A = (MS)$, then we have $B = -M^{-1}SU_{n-t}$. Hence

$$N = \begin{pmatrix} -M^{-1}S \\ I_{n-t} \end{pmatrix} U_{n-t}.$$

Thus a triangular basis is obtained from a fundamental basis by postmultiplying with an upper triangular matrix. From matrix algebra alone, it is hard to see that triangular bases can be sparser than fundamental bases. The results in this paper show that judiciously constructed, they are sparser.

Outline of this paper. In this paper we report on the design and implementation of algorithms to compute fundamental and triangular null bases. A null basis is computed by repeatedly executing an algorithm to compute a null vector. Since sparsest null bases are characterized by the greedy algorithm, a heuristic strategy of computing the basis by repeatedly finding sparse null vectors is justified.

The algorithm to compute a null vector has two phases: in the first combinatorial phase, we identify the nonzero positions in the null vector. The nonzeros in each null vector corresponds to columns of A in a circuit. In a second numeric phase, numeric values of the nonzeros are computed.

In § 2 we design a circuit algorithm that finds a dependent set from a maximum matching of A. If the matrix satisfies a nondegeneracy assumption called the weak Haar property, then this dependent set is a circuit. We also show that any circuit of A can be found from an appropriate matching. The matching theory needed to understand this paper is introduced as needed in this section.

We use the circuit algorithm to find a fundamental null basis in § 3. All circuits in a fundamental basis are computed from one fixed matching in the matrix. We report on an implementation and on our computational results.

Section 4 describes the triangular algorithm to compute triangular null bases. Here a modified circuit algorithm which chooses columns to add to a start column such that a sparse circuit is obtained is used to find circuits. The algorithm is guided in its choice of columns by a matching which is constructed simultaneously. Thus each circuit is obtained from a separate matching.

Ensuring the correctness of the triangular algorithm is a subtle issue that involves some matching theory; we prove that if a complete matching is maintained in the nonstart columns in the matrix, correctness can be assured. This "outer" matching is distinct from the matching from which a circuit is obtained.

In § 5 an implementation of the triangular algorithm is described and our results are discussed. We compare our results with the results of Gilbert and Heath (1987).

In § 6 we list some results on sparse orthogonal null bases we have obtained in Coleman and Pothen (1986b), summarize our work, and make some additional remarks.

By convention a term is in *slanted* font when it is being defined. We also denote the set operations $A \cup \{b\}$, $A \backslash \{b\}$, and $A \cup \{b\} \backslash \{c\}$ by $A + b$, $A - b$, and $A + b - c$, respectively.

2. A circuit algorithm. A sparse null vector is computed in two phases. In the first phase, a circuit is identified from a matching in the matrix. In the second phase, the nonzero coefficients of a null vector are computed by solving a system of equations. We proceed to introduce the matching theory needed to design a circuit algorithm.

The *bipartite graph* $G(A)$ of the matrix A has a row vertex corresponding to each row of A, and a column vertex corresponding to each column of A. An edge joins a row vertex to a column vertex if and only if the corresponding matrix element is nonzero. The structure of a matrix and its bipartite graph are shown in Fig. 2.1. The symbols "×", and "⊗" denote nonzeros; the rest are zeros.

A *matching* in A is a set of nonzeros of A such that no two elements in the set are chosen from the same column or the same row. A matching of A corresponds in $G(A)$ to a set of edges no two of which are incident on a common vertex. A matching in the matrix of Fig. 2.1 is shown by circled nonzero elements; in the bipartite graph the edges in the matching are drawn with thick lines. A vertex is *matched* if it is an endpoint of an edge in a matching. A vertex that is not matched is *unmatched*. The matching $\mathcal{M}$ in the figure has maximum cardinality, and hence is a *maximum matching* of A. The *match-*

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \left(\begin{array}{cccccc} \times & \times & 0 & \otimes & 0 & 0 \\ \times & 0 & \otimes & \times & 0 & \times \\ 0 & 0 & 0 & 0 & \otimes & \times \end{array}\right) \end{array}$$

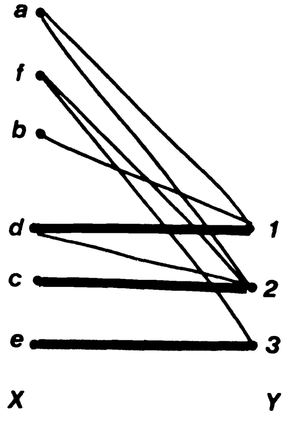


FIG. 2.1. *The structure of a matrix, its bipartite graph, and a matching.*

ing number, $m(A)$, is the cardinality of a maximum matching of A. A matching in which all the rows are matched is a *complete* (row-perfect) matching in A. A matching in which all rows and columns are matched is a *perfect* matching.

There exist several polynomial time algorithms to find maximum matchings in bipartite graphs. Let τ denote the number of nonzeros in A. The theoretically fastest known algorithm is due to Hopcroft and Karp (1973), and has time complexity $O(t^{1/2}\tau)$. Duff (1981) prefers an $O(t\tau)$ algorithm which he finds is faster in practice. Good discussions of matching algorithms may be found in Papadimitriou and Steiglitz (1982), and Lawler (1976).

The following two propositions are well known; Bondy and Murty (1976) have a proof of the first. The matrix A has the *Hall Property* (HP) if every subset of its rows has nonzeros in at least as many columns.

PROPOSITION 2.1 (Philip Hall). *A has a complete matching if and only if it has the Hall property.*

PROPOSITION 2.2. *The matching number of a matrix is greater than or equal to its rank.*

From Proposition 2.2, a matrix with rank t has a complete matching. A stronger condition on A is the Strong Hall Property (SHP). The matrix A has the *Strong Hall property* if every subset of $0 < k < n$ rows has nonzeros in at least $k + 1$ columns. (Thus when $t < n$, every set of $k \leqq t$ rows has nonzeros in $k + 1$ columns, and when $t = n$, every set of $k < n$ rows has nonzeros in $k + 1$ columns.) SHP is the same property as irreducibility. The terms HP and SHP are due to Coleman, Edenbrandt and Gilbert (1986).

A complete matching $\mathcal{M}$ of A partitions the columns of A into two sets: M, the set of matched columns, and U, the set of unmatched columns. In Fig. 2.1, $M = \{c, d, e\}$ and $U = \{a, f, b\}$. We now show that for a column $u \in U$, we can construct a circuit of A containing u by an "alternating path algorithm."

A *path* in a graph is a sequence of distinct vertices $v_1, \cdots, v_k$, where (v_{i-1}, v_i) is an edge of the graph, for $1 < i \leqq k$. An *$\mathcal{M}$-alternating path* is a path whose edges are alternately chosen from the matching $\mathcal{M}$ and outside $\mathcal{M}$. In Fig. 2.1 the sequence of edges $(b, 1)$, $(1, d)$, $(d, 2)$, $(2, c)$ is an $\mathcal{M}$-alternating path in A. We say that c and d are reachable from b by $\mathcal{M}$-alternating paths, and indicate this by $b \overset{M}{\to} c$ and $b \overset{M}{\to} d$.

An *augmenting path* is an alternating path which begins and ends with unmatched vertices. By making matched edges along an augmenting path unmatched, and vice versa, the size of the matching can be increased by one.

For $u \in U$, the following algorithm constructs a dependent set $n(u)$ containing u; this is a circuit if A has the Weak Haar Property (WHP). A matrix has the *weak Haar property* if every set of columns C satisfies rank $(C) = m(C)$. This assumption ensures that $n(u)$ will be a circuit for all "general" numeric values of the columns of A. For a particular set of numeric values of the nonzeros of A, numerical cancellations may occur, in which case the set $n(u)$ will contain a circuit. (The definitions of HP, SHP and WHP are tabulated in Table 1 for easy reference.)

TABLE 1
Summary of properties.

Hall Property (HP)	every subset of k rows has nonzeros in at least k columns
Strong Hall Property (SHP)	every subset of k rows has nonzeros in at least $k + 1$ columns
Weak Haar Property (WHP)	every subset of columns C has $m(C) = \text{rank }(C)$

THE CIRCUIT ALGORITHM. Given a matrix A with WHP, a complete matching $\mathcal{M}$, and an unmatched column $u \in U$, this algorithm finds a circuit $n(u)$.

Follow all $\mathcal{M}$-alternating paths from u, adding columns visited to $n(u)$.
Thus $n(u) = u + \{v \in M : u \overset{M}{\rightarrow} v\}$.

From Fig. 2.1 it is easy to see that $n(a) = \{a, d, c\}$, $n(b) = \{b, d, c\}$, and $n(f) = \{f, c, e\}$. The set $n(u)$ can be constructed in $O(\tau)$ time by a depth first search.

THEOREM 2.3. *The set of columns $n(u)$ is a circuit if A has WHP.*

Proof. Let C be the set of columns in the dependent set $n(u)$, and let C have nonzeros only in the row set R. For ease of notation, denote by B the submatrix A_{RC}.

We first show that B has SHP. Consider any subset S of k rows of B. S is matched in $\mathcal{M}$ to k columns, all of which are in C. If the unmatched column u has a nonzero in any of the rows in S, then S has nonzeros in at least $k + 1$ columns.

Suppose that u has no nonzero in S. Since rows in S are reachable from u by $\mathcal{M}$-alternating paths, there must exist a column, matched to a row outside S, with a nonzero in S. Again, S has nonzeros in at least $k + 1$ columns.

Let b be any column in B. Since B has SHP, $B - b$ has HP. By Proposition 2.1, $B - b$ has a complete matching of size $|R|$. Since A has WHP, the rank of $B - b$ is $|R|$, and so the columns in $B - b$ are independent. Since B is dependent, it follows that $n(u)$ is a circuit. □

Let $\tilde{C}$ denote the submatrix of columns in $C - u$ and rows in R, and let $\tilde{u}$ denote the components of u corresponding to rows in R. The coefficients of the null vector can be computed by solving

$$\tilde{C}x = -\tilde{u},$$

and then choosing

$$n(u)_i = \begin{cases} x_i & \text{if } i \text{ corresponds to a column in } \tilde{C}, \\ 1 & \text{if } i \text{ corresponds to } u, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that $\tilde{C}$ does not have WHP. Since it has a perfect matching by construction, it is rank deficient. Hence it may not be possible to express $\tilde{u}$ as a linear combination of columns in $\tilde{C}$. In this case, the coefficient of u in the null vector is zero, and we say that *the dependence in $n(u)$ does not involve u.* However, it is possible to choose a column m which has a nonzero coefficient in the null vector. Thus a null vector $n(m)$ with a nonzero component corresponding to m is obtained.

COROLLARY 2.4. *If $n(u)$ does not have WHP, it contains a circuit which can be identified by a numeric factorization.* □

We now prove that the converse of Theorem 2.3 is true.

THEOREM 2.5. *Every circuit of a matrix A with WHP can be constructed by the circuit algorithm from some maximum matching $\mathcal{M}$ of A.*

Proof. Let C be the set of columns in a circuit, and let R and B be as in the proof of Theorem 2.3. Denote $|C|$ by c, and distinguish any one column of B as u. We claim $B - u$ has HP.

Suppose not. Then R has a subset of k rows adjacent to fewer than k columns, for some k. The columns and rows of the submatrix B can be permuted to the structure in Fig. 2.2. The submatrix B then contains a dependent set of columns of size $c - k$, violating the minimality of C. Hence $B - u$ has HP, and by Proposition 2.1, it has a complete matching $\mathcal{M}_1$. Partition A as shown in Fig. 2.3.

$$\begin{array}{c|c|c|c|}
 & u & <k & \geq c-k \\ \hline
k & & & 0 \\ \hline
c-k-1 & & & \\ \hline
\end{array}$$

FIG. 2.2. *The submatrix B.*

$$\begin{array}{c|c|c|c|}
 & u & C & \hat{C} \\ \hline
R & & & \\ \hline
\hat{R} & 0 & 0 & \\ \hline
\end{array}$$

FIG. 2.3. *A partition of A.*

In any matching of A, the row set $\hat{R}$ can match only to the column set $\hat{C}$. Let $\mathcal{M}_2$ be a maximum matching of the submatrix $A_{\hat{R}\hat{C}}$. The required maximum matching is $\mathcal{M}_1 \cup \mathcal{M}_2$. □

3. Fundamental null bases. We now develop an algorithm to compute fundamental null bases using the circuit algorithm.

THE FUNDAMENTAL ALGORITHM.

```
1. [initialize] Let $N$ be the empty set;
2. [match]
     Find a complete matching $\mathcal{M}$ of $A$;
     partition the columns: $A = (MU)$;
3. [construct basis]
   for each $u \in U \rightarrow$
     construct $n(u)$ by the circuit algorithm;
     solve for the coefficients in $n(u)$;
     Augment the null basis $N$ with the computed null vector;
     rof
```

When A has WHP, by Theorem 2.3, each set $n(u)$ is a circuit. Further, since an unmatched column u is contained only in the circuit $n(u)$ by construction, N is a fundamental null basis. Thus the algorithm is correct in this case.

Step 3 of the fundamental algorithm can be modified to reduce its complexity. With the partition $A = (MU)$, the fundamental basis has the structure in equation (1.1), where $B \equiv -M^{-1}U$. Thus when M has full rank, the coefficients of each null vector $n(u)$ can be obtained by solving a system of the form $Mx = -u$. Hence we do not need to identify columns in $n(u)$ by following alternating paths. Also, the $(n - t)$ matrix factorizations can be replaced by one.

This observation also shows that when A does not have WHP, the algorithm can fail to compute all the $(n - t)$ linearly independent null vectors in a basis. Corresponding to each fundamental basis, there is an associated partition of the columns of A into M and U. Since B satisfies the equation $MB = -U$, when M does not have full rank, it may not be possible to express a column u as a linear combination of the columns in M.

Thus when A does not have WHP, a fundamental basis can be computed only when M has full rank. Hence we choose M by a matching, but ensure that M has full rank when we factor it to compute the null vectors. If it is rank-deficient, we reject the dependent columns in M from the matching, and find a new maximum matching. This strategy will ensure correctness; and will always succeed when A has full row rank.

For some of the problems reported here, the submatrix M chosen by a matching was indeed rank deficient. The number of dependent columns was almost always equal to one or two out of a few hundred columns; the largest we observed was five.

Details of implementation. Since computing a sparsest fundamental null basis is NP-hard, heuristic strategies have to be employed to find sparse bases. Our strategy is to assign costs to the columns of A, and to choose a column of minimum cost to match to a row.

The cost of a column c is the number of nonzeros in it. To justify this, observe that for a matrix with WHP, the number of nonzero rows in a circuit is one smaller than its number of columns. Thus a sparse circuit has few nonzero rows. Hence the cost of a column c indicates that any circuit containing c must have at least this number of additional matched columns in it. The cost of a matching is the sum of the costs of the matched columns. Ties were broken in favor of lower numbered columns.

Our weighted maximum matching routine is derived from MC21A, Duff's algorithm for finding a maximum matching in a matrix (Duff (1977)). A maximum matching is obtained by matching the rows one by one. At one step of the matching algorithm, we search for a column to match to an unmatched row. From the given row, a depth first search is performed through alternating paths to visit every unmatched column that could be reached by such a path. The cheapest of these columns is chosen.

The solution of the linear systems to compute the coefficients of the null vector is accomplished by using the LUSOL package of Gill, Murray, Saunders and Wright (1986). This package is presently a part of MINOS (Murtagh and Saunders (1983)). The LUSOL routines draw on the work of Reid (1976), (1982) on sparse LU factorizations of unsymmetric matrices. Gaussian elimination with row and column pivoting is performed such that $M = LU$, where the matrix PLP^t is lower triangular, PUQ is upper triangular, and P, Q are permutation matrices. Markowitz's criterion is used to select the pivot element, subject to a bound on the size of elements in L for numerical stability. Two triangular systems are solved to compute the null vector from the factors; we call this a *solve*. Parameters in LUSOL were set at their default values.

The matching algorithm has complexity $O(t\tau)$; Duff (1981) reports an $O(t) + O(\tau)$ experimental behavior. Since the dimension of M is t, the cost of factoring it in step 3 could be $O(t^3)$, and the $(n - t)$ solves could cost $O((n - t)t^2)$ operations. However, since M is sparse, a more realistic cost should be about $O(t^2)$ for the factorization and $O((n - t)t)$ for the solves.

Storage requirements of the algorithm are dominated by the storage required for A and the null basis N. The matrix A is stored in Sparspak column oriented data structures (George and Liu (1981)). Nonzeros and row indices are stored in column major order. For the use of matching routines, column indices of the nonzeros are stored in row major order. These require a double precision array of length τ, two integer arrays of length τ, and integer arrays of length $n + 1$ and $t + 1$. Let $\tau(M)$ denote the number of nonzeros in M; this is smaller than τ. LUSOL needs the submatrix M stored as an element list with parallel integer arrays for row and column indices. The factorization is stored in the data structure for M, for which a minimum length of $\tau(M) + 4t$ is recommended. The null basis N is stored in Sparspak column oriented data structures of length equal to the number of nonzeros in the basis.

Results. We implemented the fundamental algorithm in FORTRAN 77; our experimental code, BASIS, is structured and modular, and we believe it represents a careful and efficient implementation. The program was run on a VAX 11/780 (with floating point coprocessor) under Berkeley 4.2 Unix at Penn State's Computer Science Department. The f77 compiler was used to compile the code.

Constraint matrices from linear programming problems were used for tests, and are shown in Table 2. The first, *murty*, was taken from Murty (1983), and all the others were supplied to us by Dr. Michael Saunders. The nonzero matrix elements were stored in double precision. Our results are tabulated in Table 3. The algorithm found bases comparable in sparsity to the input matrices for all problems except *brandy*, for which there was a four fold increase in density. This seems to be caused by the restriction to fundamental bases, as will be seen in the next section.

The total time (seconds) reported includes the time needed to find a maximum matching, compute the LU factors of M, and solve for the null vectors. The matchings were found quite fast, and the relatively larger times for the problems *brandy*, *capri* and *etamacro* were caused by dependence in M which necessitated finding a second matching. This step can be speeded up if the current matching is updated instead of finding a new matching as we have done.

Both *murty* and *israel* had embedded identity matrices of dimension t; thus these null bases were anomalously easy to find. The times reported for these problems should therefore be considered low.

The time for the factorization phase was surprisingly low. This is due to the high sparsity in M as a result of the column selection strategy in the matching algorithm, and the efficient method for computing sparse factors via LUSOL. Solving for the coefficients accounted for most of the time needed by the algorithm; solving for each null vector took less than a tenth of a second, but the large number of null vectors caused the large

TABLE 2
Test problems.

Problem	Rows	Cols	Nonzeros	Density (%)
murty	12	30	56	15.6
afiro	27	51	102	7.4
adlittle	56	138	424	5.5
share2b	96	162	777	5.0
share1b	117	253	1179	4.0
beaconfd	173	295	3408	6.7
israel	174	316	2443	4.4
brandy	193	303	2202	3.8
e226	223	472	2768	2.6
capri	271	482	1896	1.5
bandm	305	472	2494	1.7
stair	356	614	4013	1.8
etamacro	400	816	2537	0.8

TABLE 3
Fundamental null bases.

	Null basis				Time (seconds)			
Problem	Rows	Cols	Nonzeros	Density (%)	Total	Match	Factor	Solve
murty	30	18	62	11.5	0.6	.02	.07	0.5
afiro	51	24	112	9.2	0.9	.05	.08	0.8
adlittle	138	82	500	4.4	3.5	.07	0.2	3.2
share2b	162	66	736	6.9	4.6	0.1	1.0	4.0
share1b	253	136	2264	6.6	13.7	0.7	1.7	11.3
beaconfd	295	122	1789	5.0	13.9	0.7	1.1	11.9
israel	316	142	2411	5.4	12.1	.02	0.4	11.6
brandy	303	110	4758	14.3	22.8	1.8	2.3	18.6
e226	472	249	3449	2.9	20.3	0.3	0.6	19.2
capri	482	211	3478	3.4	27.6	4.5	2.5	20.3
bandm	472	167	2306	2.9	19.5	0.7	0.1	17.4
stair	614	258	5378	3.4	35.7	1.2	1.2	33.2
etamacro	816	416	3929	1.2	39.1	2.8	2.5	33.8

time requirement. We conclude also that the numerical phase of the algorithm dominates the combinatorial phase in computational time required.

The numerical quality of each null vector was checked by computing the residual of the system of equations used to find a null vector. In all cases, this was below machine precision. Condition numbers were estimated for constraint matrices and null bases of the first eight problems in Table 2 by the LINPACK condition estimator. The estimates for the null bases were lower than the estimates for the constraint matrices, except for share1b; here the null basis had a condition estimate of approximately 10^6, about ten times that of the constraint matrix.

4. The triangular algorithm. We now describe the *triangular algorithm* that computes a triangular null basis by matching. The diagonal elements of the triangular basis correspond to a set of *$n - t$ start columns* in A. The algorithm computes a circuit containing each start column. Linear independence of the set of circuits follows from the structure of N.

Throughout this section we do not assume that A has WHP; hence the set $n(u)$ found from a matching will not necessarily be a circuit, but only a dependent set. The corresponding null vector may not have a nonzero coefficient corresponding to u. By a *null vector* $n(u)$ we mean a null vector with a nonzero component corresponding to u. In the description of the triangular algorithm, the remedial actions necessary in the absence of WHP are included.

THE TRIANGULAR ALGORITHM. Given a complete matching $\mathcal{M}$ in a matrix A, and a partition into matched columns M and unmatched columns U, this algorithm computes a triangular null basis N. The set S is the set of columns which have already been used as start columns.

```
$S := \emptyset$;
while $U \neq S$ do
        Choose a column $u \in U - S$;
        Construct the dependent set $n(u)$ from columns in $A - S$;
        Solve for the corresponding null vector;
```

```
if dependence involves $u$
        then { the null vector is $n(u)$}
            $S := S + u$; $N := N + n(u)$;
        else {$u$ has a zero coefficient in the null vector}
            find a column $m$ involved in the dependence;
            Let $n(m)$ be the associated null vector;
            $S := S + m$; $N := N + n(m)$;
            if $m \in M$ then {update the matching $\mathcal{M}$}
                    let $r$ be the row matched to $m$;
                    $\mathcal{M} := \mathcal{M} - (r, m)$;
                    Augment $\mathcal{M}$ by matching $r$
                    to a column in $A - S$;
                    Let $v$ be the newly matched
                    column;
                    $M := M - m + v$; $U := U +$
                    $m - v$;
                fi
        fi
od
```

Description of the algorithm. Let $S = \{s_1, \cdots, s_{i-1}\}$ be a set of start columns for which null vectors $\{n(s_1), \cdots, n(s_{i-1})\}$ have been computed. Columns in S will not be used in any of the null vectors to be computed in the future. The triangular algorithm maintains the invariant that $A - S$ has a complete matching $\mathcal{M}$. (This matching may change in the course of the algorithm.) The matching $\mathcal{M}$ partitions the columns of A into a set M of matched columns and a set U of unmatched columns, and $S \subseteq U$.

There is a great deal of freedom in how a dependent set $n(u)$ is constructed in this algorithm. As in the fundamental algorithm, we could employ the circuit algorithm to find $n(u)$ from the matching $\mathcal{M}$. Or, we could find $n(u)$ from *another* matching in $A - S$, with a view toward obtaining a sparse null vector. Indeed, we choose to do the latter.

Later in this section, we present a modified circuit algorithm that, given u, forms $n(u)$ by *choosing* columns in $A - S$ by simultaneously constructing a matching in $n(u)$. This matching is different from the complete matching $\mathcal{M}$. Thus for each start column u, a different matching in $A - S$ is constructed. This permits more intelligent choices in the column selection strategy to achieve sparsity. It is still essential to maintain a complete matching $\mathcal{M}$ in $A - S$ to ensure the correctness of the triangular algorithm.

Initially the complete matching $\mathcal{M}$ partitions the columns into M and U, and S is empty. For a column $u \in U - S$, a dependent set $n(u)$ is constructed by a matching algorithm from the columns in $A - S$. The corresponding null vector is computed as described in § 2, by a numeric factorization. If u is involved in the dependence, then N is augmented with the vector $n(u)$, and u is added to the set S. Otherwise, we can identify a column m with a nonzero coefficient in the null vector, and we obtain a null vector $n(m)$. The column m is added to the set S, and N is augmented with the vector $n(m)$.

If $m \in M$, then the row r matched to m in $\mathcal{M}$ has to be matched to another column in $A - S$ to maintain the invariant. This is accomplished by augmenting $\mathcal{M} - (r, m)$ to a complete matching, and updating the sets M and U.

It is easily seen that a triangular basis is obtained if the null vectors are arranged in the reverse order in which they are computed.

The modified circuit algorithm. We describe the algorithm used to find the dependent set $n(u)$. The modified circuit algorithm is a variant of an algorithm proposed by Gilbert and Heath (1986), based on the matching theory developed in this paper. First, we describe our version.

THE MODIFIED CIRCUIT ALGORITHM. Given a column u, this algorithm finds a dependent set $n(u)$. Here S is the set of start columns for which null vectors have been computed by the triangular algorithm, C is the set of active columns, and R is the set of active rows.

```
$C := \{u\}$;
$R := \{$rows in which $u$ has nonzeros$\}$;
while there is an unmatched row $r \in R$ do
    Find an augmenting path from an unmatched active row $r$ to an
    inactive column $c \in A - S$;
    Augment by adding $r$ and $c$ to the matching;
    $C := C + \{c\}$;
    $R := R + \{$inactive rows in which $c$ has nonzeros$\}$;
    od
```

This algorithm identifies an *active submatrix* formed from a set R of *active rows* and a set C of *active columns* such that A_{RC} is dependent. Initially, the only active column is the column u, and the active rows are the rows in which u has nonzeros. A queue of active rows is maintained by the algorithm. At each step, a column is chosen to match to a row in the queue. The column is added to the set of active columns, and inactive rows in which the column has nonzeros are made active and added to the queue. At termination the active rows are perfectly matched to the active columns (excluding u). If $n(u)$ has WHP, by Theorem 2.3, the active columns form a circuit.

Assume that in case of failure to find a column c to match to a row, the modified circuit algorithm terminates. We can prove that this will not happen; i.e., the modified circuit algorithm will not terminate without finding a dependent set $n(u)$.

THEOREM 4.1. *Let $A - S$ have a complete matching $\mathcal{M}$ which partitions A into M and U, with $S \subset U$. Then for a column $u \in U - S$, the modified circuit algorithm will find a dependent set $n(u)$ containing columns from $A - S$.*

Proof. If the algorithm succeeds in finding a column to match, the size of C increases by one in each iteration of the **while** loop. If it fails, the algorithm terminates. In either case, termination is assured.

If all rows in R are matched at termination, then they are matched to columns in $C - u$, and from Theorem 2.3, the columns in C form a dependent set. Hence assume that the algorithm terminates with a matching $\mathcal{M}_1$ which matches columns in $C - u$ to a subset of rows in R. Let $R_u \subseteq R$ denote the set of unmatched rows which cannot be matched by finding augmenting paths.

Let $\mathcal{M}_2$ denote the edges in the complete matching $\mathcal{M}$ incident on the rows in R. By Theorem 4.1 of Lawler (1976, Chap. 5) (also Gale and Hoffman (1982)), it is possible to find a matching $\mathcal{M}_3$ from $\mathcal{M}_1$ and $\mathcal{M}_2$ in which all rows in R and all columns in $C - u$ are matched. Thus it is possible to augment the matching $\mathcal{M}_1$ by matching rows in R_u, and this is a contradiction. □

Correctness of the triangular algorithm. We establish the correctness of triangular algorithm next. In view of Theorem 4.1, we need prove only that it is possible to maintain the invariant of the algorithm, after a start column is chosen and a null vector is computed.

THEOREM 4.2. *Let $1 \leqq i \leqq n - t$, and $S = \{s_1, \cdots, s_{i-1}\}$ be a set of start columns for which null vectors $n(s_1), \cdots, n(s_{i-1})$ have been computed. Let $A - S$ have a complete* *matching $\mathcal{M}$. There exists a column $s_i \in A - S$ from which the triangular algorithm computes a null vector $n(s_i)$ from the columns in $A - S$, such that $A - (S \cup s_i)$ has a complete matching.*

Proof. The modified circuit algorithm finds a dependent set $n(u)$ from any column u which is unmatched in $\mathcal{M}$. If the dependence involves u, then $s_i := u$, and the result is true. Otherwise, let m be a column in $n(u)$ involved in the dependence. Now $s_i := m$, and $n(m)$ is the null vector obtained.

Let M denote the columns matched in $\mathcal{M}$, and U the unmatched columns. Denote $S \cup s_i$ by $\hat{S}$. To maintain the invariant, there are two cases to consider.

Case 1. $m \in U - S$. Then M remains a completely matched set of columns in $A - \hat{S}$.

Case 2. $m \in M$. In the matching found by the modified circuit algorithm, the set of columns $n(u) - u$ is perfectly matched to the set of rows in which $n(u)$ has nonzeros. Call this set of rows R. (See Fig. 4.1.) Then from the proof of Theorem 2.5, the set of columns $n(u) - m$ can be perfectly matched to R. Since $n(u) - m \subseteq A - \hat{S}$, it is possible to match all rows in R to columns in the latter set.

Let $\hat{R}$ denote the rest of the rows of A. Then columns in $n(u)$ have zeros in the rows in $\hat{R}$, and hence a row in $\hat{R}$ cannot, in any matching, match to any of these columns. Thus columns matched in $\mathcal{M}$ to rows in $\hat{R}$ are disjoint from columns in $n(u) - m$. It follows that all rows in this set can be matched to columns in $A - \hat{S}$. Hence $A - \hat{S}$ has a complete matching.

Let r be the row matched in $\mathcal{M}$ to m. It follows from the correctness of the maximum matching algorithm that $\mathcal{M} - (r, m)$ can be augmented to a complete matching by matching r to a column in $A - \hat{S}$. □

Gilbert and Heath use a version of the modified circuit algorithm as a component in their matching algorithm (GHM) to find triangular bases. The major difference is that they maintain a QR factorization of the active submatrix as each column is being added. This helps them terminate the algorithm when a numerical dependence is detected, even when all active rows have not been matched.

We find a dependent set by matching methods alone, and then solve for the coefficients in the null vector by a sparse LU factorization of the submatrix A_{RC}. There are two advantages to such a choice. Since all the columns and rows in the dependent set are known, the row and column ordering strategies of a sparse matrix factorization can be used to keep the LU factors sparse. More important, a sparse matrix storage scheme can be used to store the nonzeros in the factors, thus keeping intermediate storage needed low.

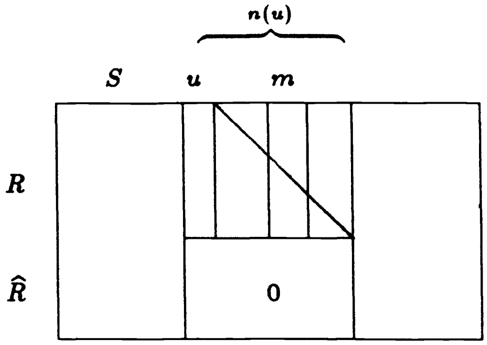


FIG. 4.1. *Proof of Theorem* 4.2: *Case* 2.

The triangular algorithm differs from GHM also in the strategy to select start columns. They use an initial QR factorization of A to identify a set of dependent columns, which is designated as the set of start columns. For each start column, they computed a null vector containing it, and any dependences not involving the start column were rejected to ensure that a basis was computed.

5. Computing triangular bases.

Implementation details. The initial matching $\mathcal{M}$ in the triangular algorithm is chosen by the column weighting strategy in § 3. The column u is chosen to be a column with most nonzeros in $U - S$, since once this null vector is computed, u will not be used again. The column m is chosen to be a column in $U - S$ rather than in M, if possible, since this saves the $O(\tau)$ operations needed to update the matching. The column v is chosen to be a column with fewest nonzeros in $U - S$ that can augment the matching.

In the modified circuit algorithm, the column c is chosen to be an inactive column of minimum cost. Here, the cost of a column c is the number of nonzeros it has in the inactive rows. Our heuristic justification is that this number of additional unmatched rows are added to a circuit when c is matched to an active row. In case of ties, two different tie breaking strategies were tried: columns with fewer total nonzeros were favored in one, and in the other, columns with most total nonzeros. The first strategy worked better for almost all problems.

Each null vector is computed by the triangular algorithm from a perfect matching in a submatrix of A. Each submatrix can have t rows and $t + 1$ columns, and hence the matching could cost $O(t\tau)$. The associated factorization could cost $O(t^3)$. The actual cost for each null vector should be lower since the number of rows in each submatrix should be small due to sparsity. Storage requirements for the algorithm are similar to that of the fundamental algorithm.

Results. An implementation of our algorithm in FORTRAN 77 forms the second part of BASIS. We present our results (computed under the same conditions as for fundamental null bases) in Table 4.

The triangular null bases we obtained are consistently sparser than fundamental bases. The increase in sparsity is most spectacular for *brandy* where the triangular basis has about half the density of the latter. In all cases, the densities of the constraint matrices and triangular null bases are comparable.

The time reported is the time (in seconds) needed to find a basis, given the matrix stored in Sparspak data structures and an initial complete matching. The triangular algorithm needs about two to five times the time needed by the fundamental algorithm. This is caused primarily by the $(n - t)$ matrix factorizations needed to compute the null vectors. The size of most of the matrices to be factored is quite small since the null vectors are sparse. Also, the LUSOL routines compute the sparse factorization efficiently. This explains why the increased time requirement is not greater.

The residuals in the system of equations from which null vectors are computed were always below machine precision. Condition numbers were estimated as before for the first eight problems in Table 2. These were all small, except for share1b and brandy (cond $(N) \approx 10^6$ for both); for the former the ratio cond (N)/cond (A) was about ten, and for the latter about one hundred.

Comparison with the Gilbert–Heath algorithms. We have also run the triangular algorithm on seven test problems shown in Table 5 from Gilbert and Heath (1987). Four of these are equilibrium matrices from structural optimization, and three, lp1, lp2 and lp3 are obtained from linear programs. The bases for the structural problems have a

TABLE 4
Triangular null bases.

	Null basis				Time (seconds)
Problem	Rows	Cols	Nonzeros	Density (%)	Total
murty	30	18	62	11.5	0.4
afiro	51	24	108	8.8	0.8
adlittle	138	82	486	4.3	3.6
share2b	162	66	686	6.4	5.0
share1b	253	136	1425	4.1	19.8
beaconfd	295	122	1581	4.4	70.4
israel	316	142	2118	4.7	34.4
brandy	303	110	2535	7.6	91.8
e226	472	249	2742	2.3	57.4
capri	482	211	2850	2.8	50.3
bandm	472	167	1941	2.5	26.4
stair	614	258	5094	3.2	59.9
etamacro	816	416	3563	1.0	59.5

natural profile structure arising from the locality of the interconnections in the physical structure, while the linear programs do not.

In Table 6, we compare the results from the triangular algorithm with results from the Gilbert and Heath matching algorithm (GHM) and turnback algorithm (GHT). Their code was executed on a reasonably similar setup to ours—a VAX 11/780 running the same operating system and on the same compiler. However, we need to be cautious about attaching too much significance to small differences in running times, since the algorithms are not being compared on the same machine.

The storage reported is the intermediate storage required to obtain a null vector. Two values are given for GHM and GHT. The dense storage reported is the maximum dense matrix storage needed for the active submatrix. The profile storage reported is the maximum storage that would be needed if a profile scheme is used to store the active submatrix. The running times pertain to an implementation that uses dense storage. The times (in seconds) in the Gilbert and Heath algorithms exclude the time needed for column pre-ordering and the initial QR factorization.

For the triangular algorithm, the storage reported is the maximum number of nonzeros in the L and U factors of the dependent set. We have not included the storage required by the integer arrays needed for column and row indices of nonzeros. This is quite small, since eight bytes are needed to store the nonzeros as double precision numbers, and only two bytes are needed to keep an integer.

TABLE 5
Test problems from Gilbert and Heath (1987).

Problem	Rows	Cols	Nonzeros	Density (%)
frame2d	27	45	93	7.7
lp1	57	97	465	8.4
frame3d	72	144	304	2.9
wheel	96	120	420	3.6
wrench	112	216	490	2.0
lp2	118	225	1182	4.5
lp3	171	320	906	1.7

TABLE 6
Results on problems from Gilbert and Heath (1987).

	Gilbert and Heath					Triangular Algorithm		
				Storage				
Problem	Algorithm	Nonzeros	Time	Dense	Profile	Nonzeros	Time	Storage
frame2d	matching	76	0.63	72	25	79	0.63	16
	turnback	76	2.30	288	27			
lp1	matching	367	10.50	1680	271	351	15.3	136
	turnback	391	28.02	2550	597			
frame3d	matching	317	2.95	168	41	310	3.6	35
	turnback	452	66.12	2064	104			
wheel	matching	488	10.45	756	273	518	5.5	103
	turnback	503	17.32	1560	326			
wrench	matching	518	26.98	3782	266	588	31.8	139
	turnback	544	58.38	5112	451			
lp2	matching	1363	103.68	8160	1784	1379	39.1	680
	turnback	1531	773.15	13570	8717			
lp3	matching	1101	60.80	3540	697	1210	78.9	169
	turnback	1518	288.87	10506	849			

The storage reported is the intermediate storage required to compute a null vector. For the Gilbert and Heath algorithms, the storage is the maximum size of the storage needed for the QR factors of the active submatrix. The dense storage refers to a dense matrix storage scheme, and the profile storage, to a profile matrix storage scheme. The reported times in GHM and GHT refer to the dense scheme. For the triangular algorithm, the storage is the maximum number of nonzeros in the sparse L and U factors of the dependent set.

Conclusions. The following conclusions may be drawn. Within the context of structural analysis problems, the use of turnback with profile storage scheme is justified. Because of the profile structure inherent in these null bases, the intermediate storage required is not prohibitive.

The two matching algorithms, GHM and the triangular algorithm, require smaller running times and less intermediate storage than turnback. The differences are greater for the linear programs, but this observation is true even for the structural problems. For instance, on lp2, GHT requires about twenty times the running time of the triangular algorithm, and more than twelve times the intermediate storage of the latter. We can conclude that a matching algorithm should be preferred over turnback for computing null bases of general sparse matrices.

Comparisons between GHM and the triangular algorithm are more difficult to make with the available data. Use of a profile scheme is essential to keep the intermediate storage from being prohibitive in GHM. However, the running times reported by Gilbert and Heath are for the dense storage scheme.

Both the algorithms compute fairly sparse null bases. The intermediate storage required by the GHM profile algorithm is of the same order as the storage required by the triangular algorithm. It is likely that the GHM profile algorithm will be a practical algorithm to compute sparse null bases.

We believe our results show clearly that the triangular algorithm is an attractive algorithm for large sparse null basis computations because of its low running times, small storage requirements, and the high sparsity achieved in the null bases.

6. Conclusions.

Sparse orthogonal null bases. In this section, we summarize some of our work on orthogonal null bases that is not included here.

We have shown that the circuit algorithm can be used to compute orthogonal null bases. By making use of the Dulmage–Mendelsohn decomposition of the square matched submatrix, we have also provided some theoretical evidence that such bases are unlikely to be sparse.

The computation of sparse orthogonal bases is further complicated by the fact that a greedy strategy may backfire; Pothen (1984) gives a counterexample to show that not choosing a sparsest null vector at a step can lead to a sparser orthogonal basis. In the nonorthogonal situation, a sparsest basis is always obtained by a greedy strategy.

We have designed algorithms to compute sparsest orthogonal bases for two special cases: a row vector of n elements, and a $t \times n$ dense matrix. For the vector, the sparsest basis has $n\lfloor \log_2 n \rfloor$ nonzeros, and for the matrix $nt\lfloor \log_2 n/t \rfloor$ nonzeros. These bases are computed by a recursive divide and conquer strategy. Proving that these bases are sparsest involves the solution of an interesting recurrence

$$f(n) = \min_{1 \leq k \leq n-1} f(k) + f(n-k) + n,$$

with $f(1) = 1$, and $f(2) = 4$. The reader can find the details in Pothen (1984) and Coleman and Pothen (1986b).

This algorithm has close connections with an algorithm to compute the orthogonal factorization on a distributed memory multiprocessor (Pothen, Jha and Vemulapati (1987)).

Summary. We have shown that matchings can be used to identify dependent sets of columns in a matrix, and thereby nonzeros in a null vector. These dependent sets are formed by choosing a start column and adding columns to it, one by one. Matchings help us in making good choices for columns to add so that a sparse null vector is obtained. This is accomplished by weighting columns and choosing a column of minimum weight to match to a row.

The resulting algorithms to compute null vectors have two phases: a combinatorial phase, in which dependent sets are identified, and a numeric phase, in which the coefficients of the null vector are computed. The time required for the second phase clearly dominates that of the first.

We have also focused attention on the structures of the null bases we construct. To ensure linear independence of the computed null vectors, the null bases are restricted to be triangular or fundamental.

To compute a fundamental basis, we need to ensure that the matched submatrix has full row rank. Only one sparse LU factorization of the matched submatrix and $(n - t)$ solves are needed to compute the basis. However, since all the null vectors are computed from a fixed initial matching, it is difficult to assign weights "globally" to columns in an intelligent way.

For each null vector in a triangular basis, we need to find a perfect matching in a submatrix, compute its LU factors, and perform a solve. In this case, a more intelligent dynamic column weighting strategy (of the modified circuit algorithm) is possible to ensure sparsity in each null vector.

Our computational results in Tables 3 and 4 demonstrate that both the fundamental and triangular algorithms succeed in computing sparse null bases, and require low running times and small intermediate storage. The triangular algorithm finds sparser bases at the expense of greater running times.

In Table 6, we compare the triangular algorithm with a turnback algorithm. The former identifies columns in a dependent set combinatorially, while the latter uses an orthogonal factorization. Consequently, the running times of the former are substantially lower. The triangular algorithm chooses columns to add to a dependent set by means of a combinatorial criterion to keep the set sparse, while turnback uses a numbering of the columns. Hence the triangular algorithm finds sparser bases. Finally, the triangular algorithm can use the column and row ordering schemes of a sparse LU factorization routine to keep intermediate storage low. The turnback algorithm cannot do so, since the columns in a dependent set are unknown before the completion of the orthogonal factorization.

For structural analysis problems with a natural profile structure in the bases, use of the turnback algorithm is justified. But for general sparse matrices, our results show that a combinatorial phase is essential to keep running times and storage low.

The conditioning of the null bases seems to depend on the conditions of the constraint matrices. When the latter were well-conditioned, the null bases were well-conditioned also, and there was less than a ten fold increase in estimated condition numbers. For one of the badly conditioned problems, the condition number increased a hundred fold. Developing algorithms that can control the conditioning of the null bases is an important open problem.

Additional remarks. Throughout this paper, we have assumed A has full row rank. The triangular algorithm can be modified to work in the rank deficient situation also. This can be done by rejecting unmatched rows in a maximum matching, since these are structurally dependent rows. The algorithm can then proceed until all unmatched columns have been used to compute null vectors. Then an LU factorization of M, the completely matched submatrix, can be used to identify the rest of the dependent columns.

Both the triangular and fundamental algorithms compute null vectors in a dependent set by computing LU factorizations; the LUSOL routines that do this have to decide when a column should be declared dependent in the factorization. This is a rather difficult numerical problem. Thus computing null bases is not immune from the difficulties associated with numerical rank determination.

The model of a circuit used by our algorithms is a submatrix with a complete matching which has one nonzero column more than its number of nonzero rows. This is appropriate when the matrix elements in A are "reasonably" random; hence, most circuits satisfy the weak Haar property.

When A is the vertex-edge incidence matrix of a directed graph, a circuit corresponds to a cycle in the graph, and a basis for the cycle space forms a null basis of A. A cycle has an equal number of nonzero columns and rows of the vertex-edge incidence matrix, unlike the circuits in this paper. Our algorithms will work correctly in this situation; however, sparser cycle bases could probably be obtained by a model that exploits this additional "structure".

Like vertex-edge incidence matrices of graphs, equilibrium matrices from structural analysis have additional structure. Most circuits have the number of nonzero rows greater than or equal to the number of nonzero columns. By considering "equilibrium graphs," bipartite graphs of equilibrium matrices, it is possible to exploit the structure in these problems, and to model circuits more accurately (Pothen (1986)). This approach yields a new algorithm to compute null bases for equilibrium matrices. This equilibrium graph algorithm succeeds in finding sparser bases faster than the triangular algorithm. In some cases even sparsest null bases can be characterized.

Computing sparse cycle bases is important in solving nonlinear programs with network constraints (Dembo (1983)). Little is known about computing sparsest cycle bases. Deo, Prabhu and Krishnamoorthy (1982) show that it is NP-complete to find sparsest fundamental cycle bases, and have designed heuristic algorithms (with implementations in Pascal) to find sparse fundamental bases.

Acknowledgments. Our thanks to John Gilbert and Mike Heath for sharing their experiences in designing their null basis code which helped us with our implementation; to Mike Saunders for sending the LUSOL package and the lp test problems to us and for being available with advice; and to the referees for their valuable suggestions to our paper.

Note added in proof. An $O(tt^3)$ algorithm has been designed to find the sparsest cycle basis of a graph by J. D. Horton (1987).

REFERENCES

M. W. Berry, M. T. Heath, I. Kaneko, M. Lawo, R. J. Plemmons and R. C. Ward (1985), *An algorithm to compute a sparse basis of the null space*, Numer. Math., 47, pp. 483–504.

M. W. Berry and R. J. Plemmons (1985), *Computing a banded basis of the null space on the Denelcor HEP multiprocessor*, Proc. AMS/SIAM Summer Conference on Linear Algebra in Systems Theory, AMS Series on Contemp. Math.

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

Thomas F. Coleman (1986), *On chordal preconditioners for large scale optimization*, Tech. Report 86-762, Computer Science Dept., Cornell Univ., Ithaca, NY.

Thomas F. Coleman, Anders Edenbrandt and John R. Gilbert (1986), *Predicting fill for sparse orthogonal factorization*, J. Assoc. Comput. Mach., 33, pp. 517–532.

Thomas F. Coleman and Jorge J. Moré (1983), *Estimation of sparse Jacobian matrices and graph coloring problems*, SIAM J. Numer. Anal., 20, pp. 187–209.

——— (1984), *Estimation of sparse Hessian matrices and graph coloring problems*, Math. Programming, 28, pp. 243–270.

Thomas F. Coleman and Alex Pothen (1986a), *The null space problem* I: *complexity*, this Journal, 7, pp. 527–537.

——— (1986b), *The null space problem* II: *algorithms*, Tech. Report 86-09, Computer Science, The Pennsylvania State University, University Park, PA; Tech. Report 86-747, Cornell Univ., Ithaca, NY.

Ron S. Dembo (1983), *A primal truncated Newton algorithm with application to large-scale network optimization*, Working Paper B-72, Yale School of Organization and Management, New Haven, CT.

Narsingh Deo, G. M. Prabhu and M. S. Krishnamoorthy (1982), *Algorithms for generating fundamental cycles in a graph*, ACM Trans. Math. Software, 8, pp. 26–42.

I. S. Duff (1977), MA28—*A set of Fortran subroutines for sparse unsymmetric linear equations*, AERE Report R8730, Harwell, England.

——— (1981), *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7, pp. 315–330.

David Gale and Alan J. Hoffman (1982), *Two remarks on the Mendelson–Dulmage theorem*, Ann. Discrete Math., 15, pp. 171–177.

Alan George and Joseph W. Liu (1981), *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ.

John R. Gilbert and Michael T. Heath (1987), *Computing a sparse basis for the null space*, this Journal, 8 (1987), pp. 446–459.

P. E. Gill, W. Murray and M. H. Wright (1981), *Practical Optimization*, Academic Press, New York.

P. E. Gill, W. Murray, M. A. Saunders and M. H. Wright (1986), *Maintaining LU factors of a general sparse matrix*, Tech. Report Systems Optimization Lab., Stanford Univ., Stanford, CA.

D. Goldfarb and S. Mehrotra (1985), *A relaxed version of Karmarkar's method*, Tech. Report, Dept. I.E. & O.R., Columbia Univ., New York.

J. E. Hopcroft and R. M. Karp (1973), *A $n^{2.5}$ algorithm for maximum matchings in bipartite graphs*, SIAM J. Comput., 2, pp. 225–231.

J. D. Horton, *A polynomial time algorithm to find the shortest cycle basis of a graph*, SIAM J. Comput., 16 (1987), pp. 358–366.

I. Kaneko, M. Lawo and G. Thierauf (1982), *On computational procedures for the force method*, Inter. J. Numer. Methods Engrg., 18, pp. 1469–1495.

Eugene L. Lawler (1976), *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York.

Bruce A. Murtagh and Michael A. Saunders (1983), *MINOS* 5.0 *user's guide*, Tech. Report SOL 83-20, Systems Optimization Lab., Stanford Univ., Stanford, CA.

Katta G. Murty (1983), *Linear Programming*, John Wiley, New York.

Christos H. Papadimitriou and Kenneth Steiglitz (1982), *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, N.J.

Alex Pothen (1984), *Sparse null bases and marriage theorems*, Ph.D. thesis, Cornell Univ., Ithaca, New York.

——— (1986), *Equilibrium graphs in structural optimization*, Tech. Report 86-22, Computer Science Dept., Penn. State Univ., University Park, PA.

Alex Pothen, Somesh Jha and Udaya Vemulapati (1987), *Orthogonal factorization on a distributed memory multiprocessor*, Hypercube Multiprocessors 1987, Michael T. Heath, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987, pp. 587–596.

J. K. Reid (1976), *Fortran subroutines for handling sparse linear programming bases*, AERE Report R8269, Harwell, England.

——— (1982), *A sparsity-exploiting variant of the Bartels–Golub decomposition for linear programming bases*, Math. Programming, 24, pp. 55–69.

D. F. Shanno and R. E. Marsten (1985), *On implementing Karmarkar's method*, Working Paper 85-01, Dept. Man. Info. Systems, Univ. Arizona, Tempe, AZ.

M. N. Thapa (1984), *Optimization of unconstrained functions with sparse Hessian matrices*, Math. Programming, 29, pp. 156–186.

A. Topcu (1979), *A contribution to the systematic analysis of finite element structures through the force method*, Ph.D. thesis, Univ. of Essen, Essen, Germany. (In German.)

P. Wolfe (1962), *The reduced gradient method*, The RAND Corporation, unpublished.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

004

EMBEDDINGS OF ULTRAMETRIC SPACES IN FINITE DIMENSIONAL STRUCTURES*

MICHAEL ASCHBACHER†, PIERRE BALDI‡, ERIC B. BAUM§ AND RICHARD M. WILSON¶

Abstract. Motivated by recent advances in theoretical physics and combinatorial optimization, we study the problem of embedding ultrametric spaces into finite dimensional structures: finite sets, Euclidean spaces $\mathbb{R}^n$, Euclidean sphere S^n, and n-dimensional hypercube with Hamming distance. We give conditions and constructions of embeddings and show a general upper bound of $n + 1$ on the cardinality of the ultrametric set. We also give an upper bound on the cardinality of quasi-ultrametric sets.

1.1. Introduction.

DEFINITION 1.1. Let (X, d) be a metric space; that is, X is a set and $d:X \times X \twoheadrightarrow \mathbb{R}^+$ is a distance function. The distance d is said to be ultrametric or non-Archimedean if it satisfies:

$$d(x,z) \leqq \max (d(x,y), d(y,z)). \tag{1}$$

Equivalently, every triangle is isosceles with the third side shorter or equal to the other two. Condition (1) implies immediately that for any two balls of radius R:

$$B(x,R) \cap B(y,R) \neq \varnothing \quad \text{implies } B(x,R) = B(y,R). \tag{2}$$

An important class of ultrametric spaces is obtained from non-Archimedean valuations over fields. For instance the p-adic valuation $|\quad|_p$ over the p-adic field $\mathbf{Q}_p$ satisfies $|x+y|_p \leqq \max (|x|_p, |y|_p)$ and the corresponding distance $d(x, y) = |x - y|_p$ is ultrametric.

Discrete ultrametric spaces are known to have a hierarchical tree-like organization and have been used for instance in taxonomy [4]. Recent advances in theoretical physics and combinatorial optimization seem to be based on the discovery of some underlying non-Archimedean structure. In the replica symmetry breaking model for the Sherrington–Kirkpatrick spin glass the geometry of the space of equilibrium states has been characterized by a hierarchic ultrametric structure [6]. A model for ultrametric information storage has been proposed in [8]. In [5] computer evidence is presented of an ultrametric organization of the 2-opt tours and the 3-opt tours in the travelling salesman problem. Similar ultrametric organization has been discussed in relation to graph coloring problems [3]. In all these cases bounds on the size of ultrametric structures can yield valuable information. In the case of the spin glass, a polynomial bound will have important consequences for the physical entropy. In the information storage models, capacity is crucial to practical applications and to biological modelling. In the optimization context, a polynomial bound on the number of λ-optima would be very surprising and might lead to algorithms yielding the shortest tour in polynomial time with probability 1. These applications are discussed in greater detail in [1].

* Received by the editors February 3, 1986; accepted for publication December 24, 1986.

† Department of Mathematics, California Institute of Technology, Pasadena, California 91125. The work of this author was supported in part by the National Science Foundation.

‡ Department of Mathematics, California Institute of Technology, Pasadena, California 91125.

§ Department of Chemistry, California Institute of Technology, Pasadena, California 91125. The work of this author was supported by National Science Foundation grant DMR-8314625 and by the Defense Advanced Research Projects Agency through arrangement with the National Aeronautics and Space Administration.

¶ Department of Mathematics, California Institute of Technology, Pasadena, California 91125. The work of this author was supported in part by National Science Foundation grant MCS82-17596.

We have thus been motivated to ask the following two questions: Let (E, d_e) and (X, d_X) be two metric spaces. Assume (X, d_X) is ultrametric. Then:

(1) Can we embed X in E; i.e., can we find a subset Y of E isometric to (X, d_X) for the distance induced by d_e on Y, $(Y, d_e|_Y) \cong (X, d_X)$?

(2) For a given E what is the maximal size of X for which such an embedding is possible?

We have studied these problems for the following metric spaces:

(a) Subsets of an n-elements set with the distance

$$d(X, Y) = \max(|X|, |Y|) - |X \cap Y|. \tag{3}$$

(b) Hypercube of dimension n, i.e., n-dimensional vectors of coordinates $(0, 1)$ or $(1, -1)$ with the Hamming distance d_h

(c) $E = \mathbb{R}^n$ with the Euclidean distance.

In § 2 we prove preliminary results concerning a class of matrices. In §§ 3 and 4 we prove the following basic theorem.

THEOREM 1.1. *For cases* (a), (b), *and* (c), $|X| \leqq n + 1$, *and this bound is attained.*

In § 5 we introduce trees. In § 6 we examine the general embedding problem. In § 7 we extend Theorem 1.1 to the case when almost every triangle satisfies equation (1). This extension is crucial to practical applications.

1.2. A class of matrices. Given a finite family $\mathcal{F}$ of real square matrices and $\lambda \in \mathbb{R}$ with $\lambda \neq A_{ij}$ for all $A \in \mathcal{F}$, define $B = B(\mathcal{F}, \lambda)$ to be the square matrix with blocks $A \in \mathcal{F}$ on the main diagonal of B and each entry of B not in such a block equal to λ. The blocks $A \in \mathcal{F}$ will be termed the **maximal blocks** of $B(\mathcal{F}, \lambda)$. Evidently if $|\mathcal{F}| > 1 < |\mathcal{F}'|$ and $B(\mathcal{F}, \lambda) = B(\mathcal{F}', \lambda')$, then $\lambda = \lambda'$ and $\mathcal{F} = \mathcal{F}'$. That is, the maximal blocks of $B(\mathcal{F}, \lambda)$ are uniquely determined, as is the parameter λ. Write $\mathcal{F}(B)$ and $\lambda(B)$ for these invariants.

Let $\mathcal{B}$ be the intersection of all sets $\mathcal{A}$ of square real matrices such that:

(B1) Each 1 by 1 real matrix is in $\mathcal{A}$.

(B2) If $\mathcal{F} \subseteq \mathcal{A}$ and $\lambda \in \mathbb{R}$ with $\lambda \neq A_{ij}$ for all $A \in \mathcal{F}$ and all entries A_{ij} of A, then $B(\mathcal{F}, \lambda) \in \mathcal{A}$.

The matrices in $\mathcal{B}$ will be termed **hierarchic**.

Define the **depth** of a 1 by 1 matrix to be 0, and, proceeding recursively, if $B \in \mathcal{B}$ with $|\mathcal{F}(B)| > 1$, define the depth $d(B)$ of B to be

$$d(B) = 1 + \max\{d(A) : \mathcal{A} \in \mathcal{F}(B)\}.$$

Given $B \in \mathcal{B}$, define the set Blk (B) of **blocks** of B as follows:

If B is 1 by 1 then Blk $(B) = \{B\}$. If $d(B) > 0$ define

$$\operatorname{Blk}(B) = \{B\} \cup \left(\bigcup_{A \in \mathcal{F}(B)} \operatorname{Blk}(A) \right).$$

Partially order Blk (B) by $A \leqq C$ if $A \in$ Blk (C).

Define a matrix B to be **ultrametric** if B is hierarchic and $\lambda(A) < \lambda(C)$ for all $A, C \in$ Blk (B) with $A < C$. Define B to be **dual ultrametric** if $-B$ is ultrametric.

LEMMA 2.1. *Let B be a nonzero ultrametric maxtrix with all $B_{ij} \geqq 0$. Then* det $(B) \neq 0$.

Proof. Recall that a real symmetric square matrix A is positive semidefinite if all eigenvalues of A are nonnegative reals and A is positive definite if all eigenvalues of A are positive reals. We use the following well-known elementary fact:

(2.1.1) Let A, C be positive semidefinite and D positive definite. Then $A + C$ is positive semidefinite and $A + D$ is positive definite.

Let $\lambda = \lambda(B)$ and n the size of B. Let J be the n by n matrix all of whose entries are 1. Then $B = \lambda J + A$, where A is a dual ultrametric matrix with $\lambda(A) = 0$ and $0 \neq A \geqq 0$. Observe λJ is positive semidefinite. Moreover if B is of depth 0, then as $B \neq 0$, $B = \lambda J = \lambda$ is positive definite. Hence, proceeding by induction on the depth of B, B is the sum of positive semidefinite matrices with a positive definite diagonal matrix. We conclude from (2.1.1) that B is positive definite. In particular det $(B) \neq 0$.

LEMMA 2.2. *Let B be a nonzero hierarchic matrix of size N. Then*

(1) *The rank of B is at least $N/2$.*

(2) *If $N \geqq 4$ and $\lambda(B) \neq 0$, then B has rank at least $N/2 + 1$.*

Proof. We perform certain row and column operations on B. Let $\mathscr{F}(B) = (B(1), \cdots, B(n))$ and let N_k be the size of $B(k)$. Set $m = N_1$. Let ${}^1B$ be the matrix obtained by subtracting the first row of B from all other rows. Observe ${}^1B(k)$ remains hierarchic and $\lambda({}^1B(k)) \neq 0$ for $k > 1$. Hence by induction on N,

(2.2.1) rank $({}^1B(k)) \geqq N_k/2 + 1$ if $N_k \geqq 4$.

It is easy to see that

(2.2.2) rank $({}^1B(k) \geqq 1, 1, 2$ for $N_k = 1, 2, 3$, respectively. In particular in this case rank $({}^1B(k)) \geqq N_k/2$.

Similarly (2.2.1) and (2.2.2) hold if $k = 1$. Next subtract the first row of ${}^1B(k)$ from the remaining rows of ${}^1B(k)$, for each $k > 1$. Denote the resulting matrix by ${}^2B$. Define m to be the size of the block $B(1)$ and write ${}^2D_i$ for the row vector $({}^2B_{i1}, \cdots, {}^2B_{im})$ of ${}^2B$. Let $v = B(1)_1$ be the first row of $B(1)$ and $\lambda(m)$ the row vector of length m all of whose entries are λ. Observe that ${}^2D_i = 0$ if $i > m$ and i is not the first row of some block, while ${}^2D_i = \lambda(m) - v$ whenever i is the first row of a block ${}^2B(k)$ with $k > 1$. Observe also that the entry in the upper right-hand corner of ${}^2B(k)$ is $\lambda(B(k)) - \lambda = \sigma_k \neq 0$. Thus if $N_k = 1$, we can add suitable multiples of column i through ${}^2B(k)$ to the first m columns of ${}^2B$ to insure that ${}^3D_i = 0$, where ${}^3B$ is the image of ${}^2B$ under these column operations.

In particular suppose ${}^3B_i$, $i \in I$ is a set of row vectors of ${}^3B$ and $\sum_{i \in I} a_i({}^3B_i) = 0$ is a linear dependence. Let $I(k)$ consist of those indices in I indexing rows in $B(k)$. Assume for each k with $N_k > 1$, the first row r_k of $B(k)$ is not in I. Then from the structure of ${}^3B$, $\sum_{i \in I(k)} a_i({}^3B_i) = 0$ for each k. Order the rows of $B(k)$ so that the last $N_k - 1$ rows contain a basis of the row space of ${}^1B(k)$ if ${}^1B(k)$ is singular. Thus:

(2.2.3) rank $(B) \geqq (\sum_k$ rank $({}^1B(k))) - \varepsilon$, where ε is the number of k such that $N_k > 1$ and ${}^1B(k)$ is nonsingular.

Assume $N \geqq 4$. We conclude from (2.2.1)–(2.2.3) that rank $(B) \geqq N/2$ and either rank $(B) \geqq N/2 + 1$ or $N_k = 2$ and rank $({}^3B(k)) = 1$ for all but at most one k_0 for which $N_{k_0} = 1$ or 3. Of course we may assume the latter and choose our ordering so that $R = ({}^3B_i : i \in I)$ is linearly independent of order M with $M \geqq N/2$, $N_j = 2$ for some $1 < j$, and with 1 not in I. To complete the proof, we may assume $\lambda(B) = \lambda \neq 0$, and it remains to show ${}^3B_1$ is independent of R. Let π be the projection of the row space on its last $N - m$ coordinates. Then ${}^3B_1$ has all entries λ and is in the space spanned by $I\pi$. This is not the case as the projection of ${}^3B(j)$ on the two columns through $B(j)$ does not contain (λ, λ) since rank $({}^3B(j)) = 1$.

3. Ultrametricity. Recall that an **ultrametric space** is a pair (S, d) where S is a nonempty set and d is a non-Archimedian distance function on S. Define a function $d{:}S \times S \to \mathbb{R}$ to be **dual ultrametric** if and only if for all $r, s, t \in S$, $d(s, s) > d(s, r) \geqq \min \{d(s, t), d(r, t)\} \geqq 0$. Finally define $d{:}S \times S \to \mathbb{R}$ to be **trimetric** if $d - \operatorname{diag}(d)$ is ultrametric and $d(x, x) \neq d(x, y)$ for all distinct $x, y \in S$, where $\operatorname{diag}(d) = d(x, y)$ if $y = x$ and 0 otherwise.

Let (S, d) be an ultrametric or trimetric space and define

$$\lambda(S) = \max \{d(a, b) : a, b \in S \text{ and } a \neq b\}.$$

For $a \in S$ define

$$\Delta(a) = \{s \in S : d(a, s) < \lambda(S) \text{ or } s = a\}.$$

Call $\Delta(a)$ the **neighborhood** of a.

LEMMA 3.1. *The set* $\{\Delta(a) : a \in S\}$ *of neighborhoods is a partition of S such that* $\Delta(a) = \Delta(b)$ *for all* $b \in \Delta(a)$.

Proof. Let $a \in S$ and suppose that $b \in S - \Delta(a)$. Let $c \in \Delta(a)$. Claim $\Delta(a) = \Delta(c)$. We may suppose that $c \neq a$. Then $d(a, b) = \lambda > d(a, c)$, so as S is ultrametric, $d(b, c) = \lambda$. Hence $S - \Delta(a) \subseteq S - \Delta(c)$. By symmetry, $S - \Delta(a) = S - \Delta(c)$, so indeed $\Delta(a) = \Delta(c)$.

Next if $s \in S$ then either $s \in \Delta(a)$ or $s \in S - \Delta(a)$. In the first case $S - \Delta(s) = S - \Delta(a)$ is nonempty by paragraph one. In the second, $a \in S - \Delta(s)$, which is then nonempty. So in any event $S \neq \Delta(s)$. Hence by paragraph one, $\Delta(s) = \Delta(t)$ for each $t \in \Delta(s)$. Thus the lemma is established.

Define the **depth** dep (S) of (S, d) recursively as follows: If $|S| = 1$ let dep $(S) = 0$. Otherwise dep $(S) = 1 + \max \{\text{dep}(\Delta(a)) : a \in S\}$.

Let $(S_i : 1 \leqq i \leqq m)$ be the set of neighborhoods $\Delta(a)$, $a \in S$, as in Lemma 3.1. Order S so that the members of S_i precede those of S_j for $i < j$, and proceeding recursively, so that each S_i and its subneighborhoods are ordered subject to the same constraint. The **distance matrix** of (S, d) is the square matrix $B = B(S)$ whose rows and columns are indexed by S and with $B_{st} = d(s, t)$ for each $s, t \in S$. Observe the following lemma.

LEMMA 3.2. *If* (S, d) *is trimetric or ultrametric then its distance matrix* $B(S)$ *is a hierarchic matrix with* $\lambda(S) = \lambda(B(S))$. *If* (S, d) *is dual ultrametric then* $B(S)$ *is a dual ultrametric matrix.*

Proof. The proof is immediate from Lemma 3.1 and the ordering of S.

LEMMA 3.3. *Let V be the space of n-tuples with* 0, 1 *entries, and d the standard inner product on V; that is* $d(u, v)$ *is the number of common nonzero entries in* $u, v \in V$. *Let* $n > 1$ *and* $S \subseteq V$.

(1) *If* (S, d) *is trimetric then* $|S| \leqq 2(n - 1)$.

(2) *If* (S, d) *is dual ultrametric then* $|S| \leqq n$.

Proof. Let $N = |S|$ and A the N by n matrix whose row vectors are the vectors in S. Observe that if A^T denotes the transpose of A, then $AA^T = B(S)$.

Embed V in n-dimensional Euclidean space $\mathbb{R}^n$ and regard A^T as a linear map from $\mathbb{R}^n$ into $\mathbb{R}^N$. Then the subspace U of $\mathbb{R}^n$ generated by S has dimension at least $\dim (UA^T) = \operatorname{rank}(B(S))$. So $n \geqq \operatorname{rank}(B(S))$. Hence Lemmas 2.1 and 2.2 complete the proof.

LEMMA 3.4. *Let S be a set of nonempty subsets of a finite set X of order* $n > 1$. *For* $s, t \in S$, *let* $d(s, t) = |s \cap t|$. *Then*

(1) *If (S, d) is trimetric then $|S| \leqq 2(n-1)$.*

(2) *If (S, d) is dual ultrametric then $|S| \leqq n$.*

Proof. This is equivalent to the proof of Lemma 3.3 since V is isometric with the set of all subsets of X via the map which takes a vector in V to its support.

Notice that the upper bounds in Lemmas 3.3 and 3.4 are attained. In Lemma 3.4(2) take S to be the set of subsets of X of order 1. In Lemma 3.4(1) let X be the set of vectors in an m-dimensional vector space W over the field of order 2 and let S be the set of cosets of all hyperplanes of W. Then $n = 2^m$ and $|S| = 2(2^m - 1)$. In this latter example S is of depth 2 with distances 2^{m-1}, 2^{m-2}, and 0.

Lemma 3.4(2) follows from a result of Ryser [9] when the depth of S is 1. Indeed our proof was suggested by that of Ryser.

LEMMA 3.5. *Let V be the space of n-tuples with* 0, 1 *entries and d the Hamming metric on V; that is $d(u, v)$ is the number of nonzero entries in u and v not common to u and v. Then $|S| \leqq n + 1$ for each ultrametric subset (S, d) of V.*

Proof. This is a special case of Lemma 4.1 in the next section, but the proof in this special case is a littler easier, and thus perhaps worth giving.

Let $N = |S|$ and let A be the N by n matrix whose rows are indexed by S and with $A_{sj} = 1$ or -1 when $s \in S$ has 1 or 0 as its jth entry. Observe that $AA^T = 2D - nJ$, where J is the N by N matrix with all its entries 1 and $D = nJ - B(S)$. Moreover D is dual ultrametric with $D \geqq 0$. Now arguing as in Lemma 3.3, the subspace U of $\mathbb{R}^n$ generated by S has dimension at least rank $(D) - 1$, as its image in $\mathbb{R}^N$ is spanned by the translates of the row vectors of $2D$ by the vector $(n, \cdots, n)$. Hence Lemma 2.1 completes the proof.

4. Euclidean space.

In this section V is n-dimensional Euclidean space over $\mathbb{R}$. For $u, v \in V$ let $\langle u, v\rangle = |u - v|$. We prove the following.

LEMMA 4.1. *Let S be an ultrametric subspace of V. Then $|S| \leqq n + 1$. Indeed translating to get $0 \in S$, $S - \{0\}$ is linearly independent.*

Assume S is an ultrametric subset of V of order N. Let $\lambda = \lambda(S)$. For $s \in S$, define $S(s) = S - \Delta(s)$. Thus $S(s)$ is the set of points in S on the sphere of distance λ from s, and $\Delta(s)$ is the set of points of S in the interior of that sphere.

As translation preserves the collection of ultrametric subsets of V, we may indeed take $0 \in S$. We first prove the following.

LEMMA 4.2. *$S(0)$ is linearly independent.*

Proof. Let A be the matrix of row vectors of $S(0)$. Then $AA^T = \lambda J - B(S(0))/2$. Notice AA^T is dual ultrametric. This is because $B(S(0))$ is ultrametric and each entry on the main diagonal of AA^T is greater than each entry off the main diagonal. Indeed each entry on the main diagonal is λ while entries off the main diagonal are of the form $\langle s, t\rangle < \lambda$ as $s \neq t$ and $|s| = |t| = \lambda$.

As AA^T is dual ultrametric, rank $(AA^T) = N$ by Lemma 2.1. Thus A^T is a surjective map from the subspace of V spanned by $S(0)$ onto $\mathbb{R}^N$ so as that space is of dimension $d \leqq N$, it follows that $d = N$ and $S(0)$ is linearly independent. So Lemma 4.2 is established.

LEMMA 4.3. $\Delta(0) - \{0\}$ *is linearly independent.*

Proof. Let $a \in S(0)$. Then $\Delta(0) \subseteq S(a)$, so $\{s - a : s \in \Delta(0)\}$ is linearly independent by Lemma 4.2. Hence $\Delta(a)$ has a linearly independent subset of order $|\Delta(a)| - 1$, so as $0 \in \Delta(0)$, the lemma follows.

Let $\Delta(0) = \{x_0, \cdots, x_m\}$ with $x_k = (x_{k1}, \cdots, x_{kn})$ and $x_0 = 0$. Appealing to Lemma 4.3 and replacing S by an image under some suitable orthogonal transformation of $\mathbb{R}^n$, we may assume that $x_{kj} = 0$ for $j > k$ and $x_{kk} = e_k \neq 0$. Let π_k be the projection of V onto the subspace V_k of V consisting of those vectors with 0 in the first k coordinates.

LEMMA 4.4. (1) *There exist* $r_i \in \mathbb{R}$, $1 \leqq i \leqq m$, *such that for all* $s = (s_1, \cdots, s_n) \in S(0)$, $s_i = r_i$.

(2) $S(0)\pi_m$ *is a linearly independent subset of* V_m.

Proof. We prove the analogous statements for $k \leqq m$ by induction on k. For $k = 0$ this is Lemma 4.1. Assume the result for $k - 1$. Then for $s \in S(0)$, $\lambda = |s| = \sum s_i^2 = |s - x_k| = \sum (s_i - x_{ki})^2$. So $0 = \sum (x_{ki}^2 - 2x_{ki}s_i) = D - 2e_k s_k$, where $D = e_k^2 + \sum_{i<k} (x_{ki}^2 - 2x_{ki}r_i)$. Thus (1) of Lemma 4.4 holds for k with $r_k = D/2e_k$. Moreover $|s\pi_k| = \lambda - \sum_{i \leqq k} r_i^2 = \lambda_k$, and for $s \neq t \in S(0)$, $|s\pi_k - t\pi_k| = |s - t|$, so $S(0)\pi_k$ is on the sphere of distance λ_k from 0 in V_k and $S(0)\pi_k$ is ultrametric in V_k. Therefore (2) of Lemma 4.4 holds by Lemma 4.1.

Notice that Lemma 4.4 completes the proof of Lemma 4.1 and that Lemmas 3.4, 3.5, and 4.1 complete the proof of Theorem 1.1 in cases (a), (b) and (c), respectively. Also as a simple consequence we have the following theorem.

THEOREM 4.5. *The maximal ultrametric set that can be embedded in the Euclidean sphere* S^n *has size* $n + 2$.

Proof. The sphere S^n is trivially embedded in the Euclidean space $\mathbb{R}^{n+1}$. Therefore an upper bound of $n + 2$ holds. On the other hand, the $(n + 1)$-dimensional hypercube can be embedded in S^n with the Euclidean distance via some trivial scaling. Ultrametric sets on the hypercube with Hamming distance are still ultrametric in $\mathbb{R}^n$ with Euclidean distance. Therefore the value $n + 2$ is attained.

5. Trees. We shall first consider the tree representation for ultrametric spaces. Let $T = (V, E)$ be a rooted tree with vertices V, edges E, and root α, $\alpha \in V$. We will define the leaves of T to be the monovalent vertices other than the root. Let $X = \{x_1, \cdots, x_k\}$ be the set of leaves of T. Let $w{:}E \twoheadrightarrow \mathbb{R}^+$ be a weight function defining the length of each edge. Let d_T be the corresponding metric on the tree. Assume that:

(5.1) There exists $h > 0$ such that $d_T(\alpha, x_j) = h$ for all j, $1 \leqq j \leqq k$.

h is called the height of the tree. More generally for every vertex v define the **height h(v)** of v to be the length of a minimal path which connects v to a leaf. Because of (5.1), $h(v)$ is well defined.

Define a metric space (X, d_X) by letting the distance between two leaves be the height of their first predecessor. Again (5.1) renders d_X well defined. It is easy to check that (X, d_X) is an ultrametric space. Moreover it can easily be shown by arguments like those of § 2 that every finite ultrametric space can be represented by such a tree.

The leaves can be partitioned into l sets: $B_1 \cdots B_l$ of nearest neighbors. We shall denote by b_i the cardinality of B_i and d_i the common distance of the leaves in B_i to their first predecessor. From now on any finite ultrametric space (X, d_x) will be an ultrametric tree with the previous conventions and with an ultrametric positive 0-diagonal distance matrix D.

We need to derive a few general matrix equations. Cases (a) and (b) with the $(1, -1)$ conventions yield the most simple expressions and this will suffice.

Case (a). Let $Y_1 \cdots Y_k$ be subsets of an n-elements set with distance:

$$d(Y_i, Y_j) = \max(|Y_i|, |Y_j|) - |Y_i \cap Y_j|.$$

Let A be the $k \times n$ incidence matrix and M be the $k \times k$ matrix defined by: $m_{ij} = \max(|Y_i|, |Y_j|)$. Then

$$AA^t = M - D. \tag{5.2}$$

In the special case where all the subsets have the same cardinality v, (5.2) yields

$$AA^t = vJ - D. \tag{5.3}$$

Case (b). Let $X_1, \cdots, X_k$ be k n-dimensional vectors of coordinates $(1, -1)$ with the Hamming distance d_h. Let B be the matrix having X_i as its ith row. Then

$$BB^t = nJ - 2D. \tag{5.4}$$

In these cases our two initial questions become the following: If D is a positive 0-diagonal ultrametric matrix under which conditions can we solve equations (5.2) and (5.4)? What is the maximal value for k if n is fixed? Notice that the tree for which the upper bound $n + 1$ of § 3 is attained has a very poor structure. One might wonder if much tighter upper bounds could be obtained for classes of trees with a richer branching structure. We shall prove now that this is not the case and examine the general embedding problem: Given a fixed ultrametric tree T can we embed it in one of the metric spaces of type (a), (b) or (c)?

6. General embeddings. We first discuss Case (a).

THEOREM 6.1. *Let T be an ultrametric tree with k leaves and D be the corresponding k by k matrix of distances. Assume D has integer entries. Then we can always embed T in an n-set for n large enough. More precisely: We can find an n-set and k of its subsets with fixed cardinality v such that the equation $vJ - D = AA^t$ is satisfied. Moreover, if h denotes the height of the tree, then $v = h$ and*

$$n = h + \sum_{i=1}^{l} (b_i - 1)d_i + \sum_{i=1}^{l-1} d_{ii+1}$$

where $d_{ii+1} = d(x_i, x_j)$ for $x_i \in B_i$ and $x_j \in B_{i+1}$.

Proof. Trivially it is necessary for the distance matrix D of T to have integer coefficients and since the weights are differences of distances they too are integers. Suppose now we are given a tree of height h such that all the weights $w(e)$ are integers. We shall construct recursively the n-set and its k h-element subsets by assigning to each vertex v of T a certain subset $f(v)$.

Let (a_n) be a list of variables. Let $P_1 \cdots P_k$ be any ordering of the k unique directed paths joining the root α to the leaves x_i, $1 \leqq i \leqq k$. Order the vertices of T lexicographically considering first the ordering of the paths and then the order within each path.

Step 1. $f(v_1) = f(\alpha) = \emptyset$.

Step m. Assume that $f(v_i)$ has been defined for $i \leqq m$ so that $f(v_i) \subseteq f(v_j)$ if $i \leqq j$ and v_i and v_j are on a common path. Let $\bigcup_{i=1}^{m-1} f(v_i) = \{a_0, a_1, \cdots, a_{g(m)}\}$. There exists a unique $p < m$ with v_p joined to v_m. Let w denote the weight of the corresponding edge. Then we set:

$$f(v_m) = f(v_p) \cup a_{g(m)+1}, \cdots, a_{g(m)+w}.$$

For any leaf x_i we have $|f(x_i)| = h$ since we start with $f(\alpha) = \emptyset$ and we add $w(e)$ new elements for any edge e belonging to the directed path between α and x_i. Let $Y_i = f(x_i) i = 1, \cdots, k$. Then by construction:

$$\max(|Y_i|, |Y_j|) - |Y_i \cap Y_j| = d(x_i, x_j)$$

which is the height of the common predecessor of x_i and x_j. Therefore $Y_1, \cdots, Y_k$ are k h-elements subsets of an n-set: $\cup_{i=1}^{k} f(x_i) = \cup_{v \in V} f(v)$ representing the given ultrametric tree T. Moreover by construction: $|\cup_{x \in B_i} f(x)| = h + (b_i - 1)d_i$. Therefore deleting all but one leaf from each block and proceeding by induction on k, we get:

$$n = h + \sum_{i=1}^{l} (b_i - 1)d_i + \sum_{i=1}^{l-1} d_{ii+1}.$$

We next consider Case (b).

We are given an ultrametric tree T and we are looking for an embedding into some n-dimensional hypercube. As in Case (a) it is easy to see that all the weights need to be integers. The same holds for h. But additional conditions are necessary as shown by the following simple lemma [11].

LEMMA 6.2. *Every triangle on the hypercube with Hamming distance d_h has an even perimeter.*

As a consequence, for every ultrametric isosceles triangle on the hypercube the third side cannot have odd length. It is easy to show by induction that a necessary condition for the existence of an embedding is that the tree T has one of the following two exclusive properties:

(i) All the weights are even integers.

(ii) The root α has only two adjacent vertices v_1 and v_2, $w(\alpha, v_1)$ and $w(\alpha, v_2)$ are odd, and all the other edges have even weights.

Such a tree will be called **hypercubic.** We can now state the following theorem.

THEOREM 6.3. *Let T be an ultrametric hypercubic tree with k leaves and distance matrix D. Then we can always embed T in an n-dimensional hypercube for n sufficiently large. More precisely: We can find k n-dimensional vectors $X_1, \cdots, X_k$ of coordinates $(1, -1)$ such that the equation $BB^T = nJ - 2D$ is satisfied. Moreover, if all edges have even length we can choose the k vectors in one of the hyperplanes:*

$$\sum_{i=1}^{n} x_i = c = \pm(n - h)$$

if the first two edges e_1 and e_2 have odd lengths $2a + 1$ and $2b + 1$ then we can choose the vectors corresponding to e_1 to be in one of the hyperplanes:

$$\sum_{i=1}^{n} x_i = \pm(n - 2c_1)$$

and those corresponding to e_2 in one of the hyperplanes:

$$\sum_{i=1}^{n} x_i = \pm(n - 2c_2)$$

with the same sign in both equations, where c_1, c_2 are two integers satisfying:

$$c_1 + c_2 = h$$

and

$$c_1 \geqq \frac{(h - 2a - 1)}{2}$$

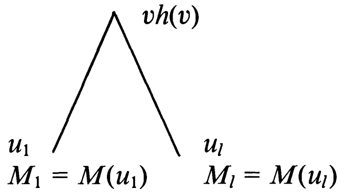
and

$$c_2 \geqq \frac{(h-2-1)}{2}.$$

Proof. For convenience we shall use (0, 1) coordinates rather than (1, −1). The **weight** of a vector will be the cardinality of its nonzero coordinates. If all row vectors in a matrix M have same weight w, we shall write $w(M) = w$: the weight of the matrix M. Proceeding recursively by height we shall now construct our embedding by attaching progressively to each vertex v of the tree a matrix $M(v)$ of weight $h(v)/2$ (except for the leaves x_i, where $w(M(x_i)) = 1$). Notice that we assume that all edges except perhaps the last two have even length. The matrix $M(\alpha)$ will provide the final embedding. The number of rows of $M(v)$ will be equal to the number of leaves attached to v.

We start by defining $M(x_i) = 1$ for every leaf x_i. Obviously, $w(M(x_i)) = 1$. Suppose we are looking now at a vertex v to which no matrix has been assigned. If $\{u_1, \cdots, u_l\} = \{u \in V: h(u) \leqq h(v) \textit{ and } (u, v) \in E\}$ and if $M(u_i)$ has been defined for $1 \leqq i \leqq l$ then we shall define a matrix $M(v)$ for the vertex v through a process called amalgamation. We shall denote: $M(v) = [M(u_1), \cdots, M(u_l)]$. We then iterate amalgamation as many times as necessary until $M(\alpha)$ is obtained. The rows of $M(\alpha)$ will represent the final vectors on the hypercube.

Definition of amalgamation. Assume we have the following situation:



Assume that M_i is $n_i \times m_i$ and $w(M_i) = h_i/2$, $i = 1, \cdots, l$ and that n_i is the number of leaves attached to u_i. Since $h_i \leqq h(v)$ we can define an integer λ_i by

$$\frac{h_i}{2} + \lambda_i = \frac{h(v)}{2} \quad \text{for } i = 1, \cdots, l.$$

Then define $M = [M_1, \cdots, M_l]$ by:

$$M = \begin{pmatrix} J_1 & 0 & & 0 & M_1 & 0 & & 0 \\ 0 & J_2 & & 0 & 0 & M_2 & & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & & J_l & 0 & 0 & & M_l \end{pmatrix}$$

where J_i is the $n_i \times \lambda_i$ matrix all of whose entries are 1. M has the following properties:

(1) $w(M) = w(M_i) + \lambda_i = h_i/2 + h(v)/2 - h_i/2 = h(v)/2$.

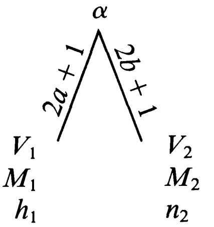
(2) M is $n \times m$ where $n = \sum_1^l n_i$ and $m = \sum_1^l (m_i + \lambda_i)$ and n is the number of leaves attached to v.

(3) The Hamming distance between any two rows i and j of M belonging to two different blocks is given by:

$$d_{ij} = w(M_i) + \lambda_i + w(M_j) + \lambda_j = h(v)$$

which is exactly the ultrametric distance between the corresponding two leaves.

If the last two edges have odd length we define the amalgamation for the corresponding two matrices in a similar way:



$$\begin{bmatrix} \overbrace{\begin{matrix} J_1 & 0 \\ 0 & J_2 \end{matrix}}^{a+b+1} & \begin{matrix} M_1 & 0 \\ 0 & M_2 \end{matrix} \end{bmatrix}$$
$$\underbrace{}_{\lambda_1}\ \underbrace{}_{\lambda_2}$$

$\lambda_1 + \lambda_2 = a + b + 1$ and $h_1 + 2a + 1 = h_2 + 2b + 1$. If all edges are even we have from property (1) $w(M(\alpha)) = h/2$. Therefore if we are using a $(1, -1)$ representation the vectors lie in the hyperplane:

$$\sum_{i=1}^{n} x_i = \frac{h}{2} - \left(n - \frac{h}{2}\right) = h - n$$

or its mirror image. If the last two edges e_1, e_2 have odd lengths then the vectors are separated into two groups of constant weight

$$w_1 = \frac{h_1}{2} + \lambda_1 \quad \text{and} \quad w_2 = \frac{h_2}{2} + \lambda_2.$$

Since $\lambda_1 + \lambda_2 = a + b + 1$ we have $w_1 + w_2 = h$ and

$$w_1 \geqq \frac{h - 2a - 1}{2}, \qquad w_2 \geqq \frac{h - 2b - 1}{2}.$$

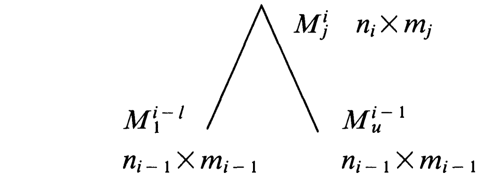
Finally using properties (1), (2) and (3), it is easy to show that the matrix $M(\alpha)$ yields the required embeddings.

Let us now consider regular trees and compute the corresponding dimension n of the hypercube.

THEOREM 6.4. *Let T be a tree such that every vertex with the exception of the leaves has a fixed number of successors u. Assume that all edges have a constant even length $2l$. Let $h = 2ml$. Then T can be embedded in an n-dimensional hypercube with*:

$$n = ul\left(\frac{u^m - 1}{u - 1}\right).$$

Proof. Consider the amalgamation step:



We have $n_i = un_{i-1} + ul$ and $n_1 = ul$. Therefore solving this recurrence relation, we get:

$$n = \sum_{i=1}^{m} lu^i = ul\left(\frac{u^m - 1}{u - 1}\right).$$

COROLLARY 6.5. *Let n be fixed. Then we can embed in the hypercube H_n an ultrametric tree with constant even valence u (with the exception of the leaves) and edges of constant even length $2l$ corresponding to an ultrametric set of size at least k:*

$$k \geqq \frac{(n+1)(u-1)}{u^2 l} + \frac{1}{u}.$$

Proof. If $n = ul(u^m - 1)/(k - 1)$ then the size is $k = u^m$. Therefore the worst case corresponds to $n = ul((k^{m+1} - 1)/(k - 1)) - 1$ for which the size is still $k = u^m$. Solving for k we get the bound of the theorem. Asymptotically this indicates that hypercubic trees with a rich branching structure can be embedded into the hypercube, the bound on the size being still of the form $O(n)$.

Finally we consider Case (c).

We are now given a tree T and want to embed it in $\mathbb{R}^n$. Surprisingly enough it is not true that every finite metric space can be embedded in $\mathbb{R}^n$ for n large enough and with the Euclidean distance.

One obvious reason is that for any three points which are not collinear, the triangle inequality must be strict. This does not yield a sufficient condition of embeddability since counterexamples can be found by slightly perturbing cases where the triangle inequality is not strict.[1]

If the finite metric space is ultrametric then the triangle inequality is obviously strict for any three distinct points.

We can now prove the following theorem.

THEOREM 6.6. *Every finite ultrametric space with rational matrix distance D can be embedded into the Euclidean space $\mathbb{R}^n$, for n large enough. Moreover, the points can be chosen in one of the hyperplanes of equation*

$$\sum_{i-1}^{n} x_i = \pm(n-h).$$

Proof. The idea is to use scaling on the given distances, obtain a new set that can be embedded into a hypercube and then go back to $\mathbb{R}^n$. Since D is assumed to have rational entries we can find a constant c such that the matrix cD has integer entries which are also multiples of 4. Construct a new matrix D' with entries $d'_{i,j}$ defined by:

$$d'_{i,j} = \frac{c^2 d^2_{i,j}}{4}.$$

Notice that by construction $d'_{i,j}$ is even. Moreover, it is easy to check that the matrix D' defines an ultrametric space. Therefore using Theorem 6.3 the corresponding set can be embedded into an n-dimensional hypercube with Hamming distance for n large enough. For points on the hypercube with 1, -1 coordinates the Hamming distance and the Euclidean distance are related by: $d_e = 2\sqrt{d_h}$. Therefore the previous construction yields, in fact, an embedding in $\mathbb{R}^n$ with distance matrix cD. To obtain the final embedding we now need only to rescale by a factor of $1/c$. Because of Theorem 6.3 the points can be

[1] The referee has pointed out that there is a necessary and sufficient condition for a finite metric space to be embeddable into Euclidean space $\mathbb{R}^n$, namely that the square of the distance be of negative type. That is, if $x_1, \cdots, x_n$ are the points, and $\lambda_1, \cdots, \lambda_n$ are arbitrary reals with $\lambda_1 + \cdots + \lambda_n = 0$, then $\sum \lambda_i \lambda_j d(x_i, x_j)^2 \leqq 0$ holds. This characterization goes back to Cayley, but was first stated in this form by Schonberg in the 1930s.

chosen in one of the hyperplanes:

$$\sum_{i=1}^{n} x_i = \pm(n-h).$$

We can now extend this to show that every ultrametric space with $n+1$ or fewer points can be embedded in $\mathbb{R}^n$.

THEOREM 6.7. *Let X be a finite ultrametric space of cardinality m with real distance matrix D. Then X can be embedded into $\mathbb{R}^{m-1}$.*

Proof. Let us denote by $D(x_1, \cdots, x_k)$ the bordered symmetric determinant of order $k+1$:

$$\begin{vmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & (d_{12})^2 & \cdots & (d_{1k})^2 \\ 1 & (d_{12})^2 & 0 & \cdots & (d_{2k})^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (d_{1k})^2 & (d_{2k})^2 & \cdots & 0 \end{vmatrix}.$$

The following theorem by Menger can be found in [2]:

A necessary and sufficient condition that a semimetric space X may be congruently embedded in the Euclidean n-dimensional space $\mathbb{R}^n$ is:

(1) For each positive integer k, $2 \leqq k \leqq n+1$, and each set of k points $x_1, \cdots, x_k$ of X, sgn $D(x_1, \cdots, x_k) = (-1)^k$ or 0.

(2) Each set of $n+2$ points of X has a vanishing bordered symmetric determinant.

Recall that if d is ultrametric so is d^2. Also note that if $n = m-1$, condition (2) is trivially satisfied.

Assume now for contradiction that we can find k points $x_1, \cdots, x_k$ of X violating condition (1). The corresponding bordered determinant therefore has sign $(-1)^{k+1}$. Yet we can slightly perturb the ultrametric matrix of distances between the points $x_1, \cdots, x_k$ so that the newly obtained matrix is still ultrametric and has rational entries. Therefore, by Theorem 6.6 and Menger's result the corresponding bordered determinant can not have sign $(-1)^{k+1}$. Since the rational approximation of the d_{ij} and hence of the d_{ij}^2 can be made with arbitrary precision, a contradiction arises by continuity.

7. Quasi-ultrametric structures. For practical applications one must study structures that are quasiultrametric in some sense. There are two natural cases: First where every triangle violates the ultrametric constraint by only a small amount; and second where almost every triangle satisfies the constraint exactly, but a small subset is allowed to violate ultrametricity. We introduce two definitions and state corresponding results.

DEFINITION 7.1. Let (E, d) be a metric space and X a subset with the induced metric. (X, d) is ε-ultrametric if and only if there exists an ultrametric subspace (Y, d) of (E, d) such that for all $x_1, x_2, x_3 \in X$ there exists $y_1, y_2, y_3 \in Y$ with $y_i \in B(x_i, \varepsilon)$.

In reference [8], $\sqrt{n}$-ultrametric structures on the hypercube are considered. Using Stirling's formula it is easy to see that $|B(x, \sqrt{n})|$ is exponential and therefore $\sqrt{n}$-ultrametric structures may be exponential in size. The same should hold for any $f(n)$-ultrametric structure on the hypercube where $f(n)$ is an increasing unbounded function of n.

In many of the applications, one considers a sequence $((X_n, d), Y_n)$, where the (X_n, d) are metric spaces and the Y_n are finite subspaces which have the property that in the limit as $n \to \infty$, almost every triangle in Y_n satisfies the ultrametric condition (1) under the induced metric. For example, in the case of infinite range spin glasses, (X_n, d)

is $\mathbb{R}^n$ under the Euclidean metric and Y_n is a set of n-vectors, the "thermodynamic equilibrium states." Describing the set Y_n is fundamental to understanding the physics of the model. It has been shown, in the "R.S.B." model (a model believed to accurately reflect the physics), that in the limit of large n, the probability that any triangle among the Y_n will satisfy the ultrametric constraint is one [6]. An important question is whether one can bound $k_n = |Y_n|$ by some polynomial in n. Such a bound will follow from Theorem 1.1 (for the Cases (a), (b) and (c)) if we can find a constant m and subspaces $U_n \subset Y_n$ such that (U_n, d) are ultrametric and $|U_n|^m \geqq k_n$, for then $(n+1)^m \geqq |U_n|^m \geqq k_n$. This motivates the following discussion.

Let (A, d) be a finite metric space of cardinality k. Let $T(A)$ be the set of triangles in A and $T'(A)$ be the subset of those triangles violating condition (1). $|T(A)| = \binom{k}{3}$. Similarly consider (A_k, d), a sequence of finite metric spaces of cardinality k, for arbitrarily large integers k.

DEFINITION 7.2. (A_k, d) is almost ultrametric iff

$$\lim_{k\to\infty} \frac{|T'(A_k)|}{|T(A_k)|} = 0.$$

(A, d) is q-almost ultrametric if

$$|T'(A)| \leqq \binom{k}{3} k^{-q} \quad \text{for } q \geqq 0.$$

We will assume in the following theorem that (A, d) is taken from one of Cases (a), (b), or (c), with A n-dimensional.

THEOREM 7.3. *If (A, d) is q-almost ultrametric, then $|A| \leqq ((3\sqrt{3}/2)n)^{2/q}$.*

Proof. This is a corollary of a theorem of J. Spencer [10] which states that the smallest set of triangles on k vertices such that there is no independent set of size l contains at least

$$\binom{k}{3}^3 \left(\frac{l-1}{2}\right)^{-2}$$

triangles. (An **independent set** is defined as a set of vertices containing no triangles.) Thus A contains an independent set S of size r, so long as

$$\binom{k}{3} k^{-q} < \frac{4}{27} \frac{k^3}{(r-1)^2}$$

which will be true for $r < (2/3\sqrt{3})k^{q/2} + 1$. Thus there is an independent set of size $(2/3\sqrt{3})k^{q/2} + 1$, and Theorem 1.1 establishes the bound on k.

Acknowledgment. We would like to thank L. E. Baum for a critical reading of the manuscript.

REFERENCES

[1] P. F. BALDI AND E. B. BAUM, *Bounds on the size of ultrametric structures*, Phys. Review Lett., 56 (1986), p. 1598.

[2] L. BLUMENTHAL, *Distance geometries*, The University of Missouri Studies, 13 (1938), pp. 56–57.

[3] J. P. BOUCHAUD AND P. LE DOUSSAL, *Ultrametricity transition in the graph coloring problem*, preprint 1985.

[4] N. JARDINE AND R. SIBSON, *Mathematical Taxonomy*, John Wiley, New York, 1971.

[5] S. KIRKPATRICK AND G. TOULOUSE, *Configuration space analysis of traveling salesman problems*, J. de Physique, 46 (1985), pp. 1277–1292.

[6] M. MÉZARD, G. PARISI, N. SOURLAS, G. TOULOUSE AND M. VIRASORO, *Replica symmetry breaking and the nature of the spin glass phase*, J. de Physique, 45 (1984), pp. 843–854.

[7] M. MÉZARD AND M. A. VIRASORO, *The microstructure of ultrametricity*, J. de Physique, 46 (1985), pp. 1293–1300.

[8] N. PARGA AND M. A. VIRASORO, *The ultrametric organization of memories in a neural network*, J. de Physique, 47 (1986), pp. 1857–1864.

[9] H. J. RYSER, *An extension of a theorem of de Bruijn and Erdös on combinatorial designs*, J. Algebra, 10 (1968), pp. 246-261.

[10] J. SPENCER, *Turán's theorem for k-graphs*, Discrete Math., 2 (1972), pp. 183–186, in Probabilistic Methods in Combinatorics, P. Erdös and J. Spencer, eds., Academic Press, New York, 1974.

[11] M. E. TYLKIN, *On the geometry of the Hamming cube*, Dokl. Akad. USSR, 134 (1960), pp. 1037–1040.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

THE EXPONENT SET OF PRIMITIVE, NEARLY REDUCIBLE MATRICES*

JIA-YU SHAO†

Abstract. In [1] and [2], R. A. Brualdi and J. A. Ross studied the exponent set of a particular class of primitive matrices—primitive, nearly reducible matrices. They obtained an upper bound on the exponent and constructed some matrices with small exponents. Ross [2] suggested considering the problem of determining the quantity $e(n)$—the least integer $e(n) \geqq 6$ such that no $n \times n$ primitive, nearly reducible matrix has exponent $e(n)$. In this paper we give a nontrivial lower bound $e(n) \geqq (n^2 - 2n + 10)/9 + 1$ by showing that every integer k with $6 \leqq k \leqq (n^2 - 2n + 10)/9$ is an exponent of some $n \times n$ primitive, nearly reducible matrix. This also extends the result ([2, § 3]) that every integer k with $6 \leqq k \leqq n + 1$ is the exponent of some $n \times n$ primitive, nearly reducible matrix.

Key words. exponent, primitive nearly reducible matrix, minimally strong directed graph

AMS(MOS) subject classifications. primary 15A48, 05C20

1. Introduction. In this paper we investigate the properties of the exponent set of a particular class of primitive matrices—primitive, nearly reducible matrices. The terminologies and notation used in this paper will basically follow those in [1] and [2]. The definitions of irreducible matrices, nearly reducible matrices, primitive matrices and their exponents $\gamma(A)$ as well as the definitions of strong digraphs, ministrong (minimally strong) digraphs, primitive digraphs and their exponents $\gamma(D)$ are as usual and can be found, for example, in [1] or [2]. For any $n \times n$ nonnegative matrix A, we define its associated digraph $D(A) = (V, E)$ to be the digraph with $V = \{1, 2, \cdots, n\}$ and $E = \{(i, j)|a_{ij} > 0\}$. Clearly $D(A)$ depends only on the zero-nonzero pattern of A. It is well known that under this correspondence of matrices and digraphs, we have the following: A irreducible $\Leftrightarrow$ $D(A)$ strong, A nearly reducible $\Leftrightarrow$ $D(A)$ ministrong, A primitive $\Leftrightarrow$ $D(A)$ primitive and in this case A and $D(A)$ have the same exponent $\gamma(A) = \gamma(D(A))$. So the exponent set of $n \times n$ primitive, nearly reducible matrices (denoted by NE_n) is just the exponent set of primitive, ministrong digraphs with n vertices and we may use the (more intuitive) graph theoretical language to formulate and prove our results.

First we recall that a digraph D is *primitive* if there exists an integer $k > 0$ such that for all ordered pairs of vertices $i, j \in V(D)$ (not necessarily distinct), there exists a walk from i to j with length k in D, and the least such k is called the exponent of D, denoted by $\gamma(D)$. The following theorem is a characterization of primitive digraphs.

THEOREM A ([5, pp. 49–50]). *A digraph D is primitive if and only if D is strong and* g.c.d$(r_1, \cdots, r_\lambda) = 1$ *where* $L(D) = \{r_1, \cdots, r_\lambda\}$ *is the set of distinct lengths of the elementary cycles of D and* g.c.d *means "the greatest common divisor."*

Next we give some more definitions.

DEFINITION 1.1. Let D be a primitive digraph, $i, j \in V(D)$. Then the (*local*) *exponent* from i to j, denoted by $\gamma(i, j)$, is the least integer γ such that there exists a walk of length m from i to j for all integers $m \geqq \gamma$.

From Definition 1.1 it is easy to see that $\gamma(D) = \max_{i,j \in V(D)} \gamma(i, j)$.

DEFINITION 1.2. Let D be a primitive digraph with the cycle length set $L(D) = \{r_1, \cdots, r_\lambda\}$. For $i, j \in V(D)$, the *relative distance* $d_{L(D)}(i, j)$ from i to j is defined to be

* Received by the editors June 2, 1986; accepted for publication December 22, 1986. This research was partially supported by the Science Foundation of Academia Sinica.

† Department of Applied Mathematics, Tongji University, Shanghai, China.

the length of the shortest walk from i to j that meets at least one cycle of each length r_i for $i = 1, 2, \cdots, \lambda$.

Let $a_1, \cdots, a_k$ be a set of distinct positive integers with g.c.d$(a_1, \cdots, a_k) = 1$. The *Frobenius number* $\phi(a_1, \cdots, a_k)$ is defined to be the least integer ϕ such that every integer $m \geqq \phi$ can be expressed in the form $m = z_1a_1 + \cdots + z_ka_k$ where $z_1, \cdots, z_k$ are nonnegative integers. A result due to Schur shows that $\phi(a_1, \cdots, a_k)$ is finite if g.c.d$(a_1, \cdots, a_k) = 1$. In the case $k = 2$, we have $\phi(a_1, a_2) = (a_1 - 1)(a_2 - 1)$.

The following basic upper bound for $\gamma(i, j)$ will be used in the proof of our main results.

THEOREM B ([4]). *Let D be a digraph, and let $L(D) = \{r_1, \cdots, r_\lambda\}$ denote the cycle length set of D. Then $\gamma(i, j) \leqq d_{L(D)}(i, j) + \phi(r_1, \cdots, r_\lambda)$ for all $i, j \in V(D)$.*

2. Some basic properties. Let $NE_n = \{m \in \mathbb{Z}^+ | m = \gamma(D)$ for some primitive, ministrong digraphs with n vertices$\}$. In this section, we investigate some basic properties of ministrong digraphs and the exponent set NE_n.

Let D be a strong digraph. A vertex x is called an *antinode* of D if both the indegree $d^-(x)$ and the outdegree $d^+(x)$ are equal to 1. A path $\pi = (x_0, x_1, \cdots, x_k)$ with $k \geqq 2$ is called a *branch* of D if $x_1, \cdots, x_{k-1}$ are antinodes of D but x_0 and x_k are not and $D \backslash \{x_1, \cdots, x_{k-1}\}$ is strong.

LEMMA 2.1. *Suppose $D = (V, E)$ is a ministrong digraph which is not an elementary cycle. H is a maximal proper strong induced subdigraph of D (maximal with respect to the inclusion of vertex sets). Then the arc set $E(D) \backslash E(H)$ forms a branch of D.*

Proof. It will suffice to prove that $E(D) \backslash E(H)$ is a path P of length $\geqq 2$ with two end vertices in $V(H)$ and all the interial vertices in $V(D) \backslash V(H)$. Since H is a proper induced subdigraph, $V(H) \neq V(D)$. Take $v \in V(D) \backslash V(H)$ such that there exists $u \in V(H)$ with $(u, v) \in E(D)$, and take a path Q from v to a vertex $w \in V(H)$ which is the nearest vertex in $V(H)$ from v. Then $P = uv + Q$ is a path of length $\geqq 2$ with two end vertices u, w in $V(H)$ and all the interial vertices in $V(D) \backslash V(H)$. Now $H' = H + P$ is a strong subdigraph of D, and since every strong subdigraph of a ministrong digraph is an induced subdigraph, H' is a strong induced subdigraph of D. By the maximality of H and the fact that $|V(H')| > |V(H)|$, it follows that $V(H') = V(D)$ and H' is the subdigraph induced by $V(D)$; thus $H' = D$ and $E(D) \backslash E(H) = E(P)$ forms a branch of D.

As a corollary of Lemma 2.1, we get the following well-known result.

COROLLARY. *Every ministrong digraph D contains an antinode.*

Proof. If D is not an elementary cycle, we use Lemma 2.1. If D is an elementary cycle, every vertex of D is an antinode.

Using the antinode of a ministrong digraph D, we can construct a new ministrong digraph $\tilde{D}$ with one more vertex which is in some sense similar to D.

LEMMA 2.2. *Let $D = (V, E)$ be a ministrong digraph, v an antinode of D with $(u_1, v) \in E$ and $(v, u_2) \in E$. Define $\tilde{D} = (\tilde{V}, \tilde{E})$ to be a new digraph with $\tilde{V} = V \dot{\cup} \{\tilde{v}\}$ and $\tilde{E} = E \dot{\cup} \{(u_1, \tilde{v}), (\tilde{v}, u_2)\}$. Then D is also ministrong.*

Proof. It is clear that $\tilde{D}$ is strong. To show $\tilde{D}$ is ministrong, take $e = (x, y) \in \tilde{E}$. If e is incident with v or $\tilde{v}$, then $\tilde{D} \backslash \{e\}$ is not strong since both v and $\tilde{v}$ are antinodes of $\tilde{D}$. If e is not incident with v and $\tilde{v}$, then $e \in E$ and there is no path from x to y in $D \backslash \{e\}$ since D is ministrong. It follows that there will be no path from x to y in $\tilde{D} \backslash \{e\}$ because any path from x to y in $\tilde{D} \backslash \{e\}$ using vertex $\tilde{v}$ (hence using arcs $(u_1, \tilde{v})$ and $(\tilde{v}, u_2)$) can be replaced by a path using (u_1, v) and (v, u_2) which avoids $\tilde{v}$. So in any case $\tilde{D} \backslash \{e\}$ is not strong and $\tilde{D}$ is a ministrong digraph.

Now we can prove the following lemma which shows that the exponent set NE_n is "ascending."

LEMMA 2.3. $NE_1 \subseteq NE_2 \subseteq \cdots \subseteq NE_n \subseteq NE_{n+1} \subseteq \cdots$.

Proof. We will prove this lemma in matrix version as we did in [4]. Suppose $A \geqq 0$ is an $n \times n$ nonnegative, nearly reducible matrix and without loss of generality, we may assume that the last row (and the last column) of A corresponds to an antinode of the associated digraph $D(A)$. Let $\tilde{A}$ be the unique $(n+1) \times (n+1)$ matrix satisfying the following two conditions:

(1) The upper left $n \times n$ principal submatrix of $\tilde{A}$ is A;

(2) The last two rows of $\tilde{A}$ are equal and the last two columns of $\tilde{A}$ are equal.

Then the associated digraph of $\tilde{A}$ is $D(\tilde{A}) = \widetilde{D(A)}$ where $\widetilde{D(A)}$ is the digraph as defined in Lemma 2.2. Now $\widetilde{D(A)}$ is ministrong since $D(A)$ is, so $\tilde{A}$ is nearly reducible. By using the same proof as in [3], we see that $\tilde{A}^k$ and $\widetilde{A^k}$ have the same zero-nonzero pattern; hence the primitivity of A implies the primitivity of $\tilde{A}$ and in this case they have the same exponent. This proves that $NE_n \subseteq NE_{n+1}$ and completes the proof of Lemma 2.3.

From Lemma 2.3 we see that in order to show that $m = \gamma(D)$ for some primitive, ministrong digraph D with n vertices, it will suffice to show that $m = \gamma(D')$ for some primitive, ministrong digraph D' with no more than n vertices. In most cases this will enable us to simplify the construction of digraphs.

3. Some exponent sets in NE_n. In this section we will construct several families of primitive, ministrong digraphs. By computing their exponents we can get several subsets of the exponent set NE_n. Then in § 4 we will combine these subsets by suitably choosing the parameters in each subset and show that their union covers the desired portion of the exponent set NE_n. The digraphs constructed here are required to possess the following features:

(1) They will have some special structures so that the exact values of their exponents can be computed.

(2) The values of their exponents should "match" (in some sense) each other.

(3) They should be ministrong.

(4) The number of vertices of the digraphs should be less than or equal to n.

The first family of digraphs are constructed in the following lemma.

LEMMA 3.1. *Let $r_1 > r_2 > \cdots > r_\lambda \geqq 2$ be integers with* g.c.d.$(r_1, \cdots, r_\lambda) = 1$ *and let $\phi = \phi(r_1, \cdots, r_\lambda)$ be the Frobenius number of $r_1, \cdots, r_\lambda$. Let t, n be integers such that $1 \leqq t \leqq r_\lambda - 1$ and $r_1 + \cdots + r_\lambda \leqq n + t(\lambda - 1)$. Then $\phi + 2r_1 - t - 1 \in NE_n$.*

Proof. Let D be the digraph in Fig. 3.1 consisting of a path $P = v_1 v_2 \cdots v_t$ of length $t-1$ and other λ paths $P_1, \cdots, P_\lambda$ from v_t to v_1 such that $P, P_1, \cdots, P_\lambda$ are pairwise internally vertex disjoint and the length of P_i is $r_i - t + 1$ for $i = 1, 2, \cdots, \lambda$. Clearly D is strong with the cycle length set $L(D) = \{r_1, \cdots, r_\lambda\}$, so the hypothesis g.c.d.$(r_1, \cdots, r_\lambda) = 1$ implies the primitivity of D. Also the hypothesis $1 \leqq t \leqq r_\lambda - 1$ implies that D is ministrong and the condition $r_1 + \cdots + r_\lambda \leqq n + (\lambda - 1)t$ means that the number of vertices of D is $m = r_1 + \cdots + r_\lambda - (\lambda - 1)t \leqq n$. Now we want to show that $\gamma(D) = \phi + 2r_1 - t - 1$. Let u, v be vertices on P_1 such that (v_t, u) and (v, v_1) are arcs of P_1, and look at $\gamma(u, v)$. From the structure of D we see that there is only one elementary path from u to v (with length $r_1 - t - 1$) and the length of any walk from u to v has the form $l = r_1 - t - 1 + a_1 r_1 + \cdots + a_\lambda r_\lambda$, where $a_1, a_2, \cdots, a_\lambda$ are nonnegative integers and $a_1 = 0 \Rightarrow a_2 = \cdots = a_\lambda = 0$. It follows that $\gamma(u, v) = \phi + 2r_1 - t - 1$. On the other hand, for any $x, y \in V(D)$, let z be the vertex on P which is nearest from x, then $d(x, z) \leqq r_1 - t$. But z (as a vertex of P) belongs to every cycle, so $d(z, y) \leqq r_1 - 1$ and thus $d_{L(D)}(x, y) \leqq d(x, z) + d(z, y) \leqq r_1 - t + r_1 - 1 = 2r_1 - t - 1$, so $\gamma(x, y) \leqq d_{L(D)}(x, y) + \phi \leqq \phi + 2r_1 - t - 1$. Now

$$\gamma(D) = \max_{x, y \in V(D)} \gamma(x, y) = \gamma(u, v) = \phi + 2r_1 - t - 1,$$

and we obtain that $\phi + 2r_1 - t - 1 \in NE_m \subseteq NE_n$. This proves Lemma 3.1.

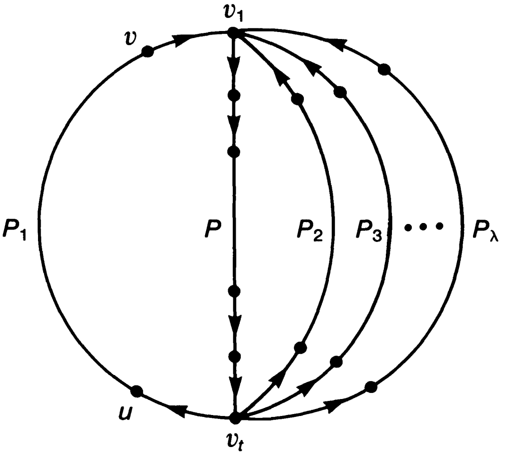


FIG. 3.1

COROLLARY 3.1. *Let r_1, r_2 be integers with $r_1 > r_2 \geqq 2$, g.c.d$(r_1, r_2) = 1$, and let $\phi = \phi(r_1, r_2)$. Suppose that $r_1 + r_2 \leqq n + 1$, then $k \in NE_n$ for all $\phi + 2r_1 - r_2 \leqq k \leqq \phi + 2r_1 - 2$.*

Proof. Take $\lambda = 2$ and $t = 1, 2, \cdots, r_2 - 1$ as in Lemma 3.1.

COROLLARY 3.2. *Let a, b be integers with $a > b \geqq 2$, g.c.d$(a, b) = 1$ and $\phi = \phi(a, b)$. If $2a + 2 \leqq n$, then $\phi + 2a + b \in NE_n$.*

Proof. Take $\lambda = 3$, $r_1 = a + b$, $r_2 = a$, $r_3 = b$ and $t = b - 1$ as in Lemma 3.1. Notice that $a + b$ is already a nonnegative integral linear combination of a and b, so $\phi(a + b, a, b) = \phi(a, b)$. Also

$$a \leqq \frac{n-2}{2} \Rightarrow r_1 + r_2 + r_3 = 2a + 2b \leqq n + 2(b-1) = n + t(\lambda - 1).$$

The hypothesis of Lemma 3.1 is satisfied and the result follows.

LEMMA 3.2. *Let a, b be integers with $a \geqq b - 1 \geqq 1$, g.c.d$(a, b) = 1$ and $\phi = \phi(a, b)$. Let t, n be integers such that $1 \leqq t \leqq a - 1$ and $2a + b \leqq n + t + 1$, then*

$$\phi + 3a - t - 1 \in NE_n.$$

Proof. Let D be the digraph constructed in Fig. 3.2 in which there are two cycles C_1, C_2 of length a with a common path $P = v_1 \cdots v_t$ and another cycle $C_3 = u_1 u_2 \cdots u_b u_1$ of length b such that $V(C_1) \cap V(C_3) = \emptyset$ and $V(C_2) \cap V(C_3) = \{u_1\}$. D contains $m = 2a - t + b - 1 \leqq n$ vertices. Clearly D is strong and since $t \leqq a - 1$ and $b \geqq 2$, D is ministrong. D is also primitive by the hypothesis g.c.d$(a, b) = 1$. Now we compute the exponent $\gamma(D)$. Let $u, v \in V(C_1)$ such that both (v_t, u) and (v, v_1) are arcs of C_1 and look at $\gamma(u, v)$. From the structure of D we see that there is only one elementary path from u to v (with length $a - t - 1$) and the length of any walk from u to v has the form $l = a - t - 1 + ha$ or $l = a - t - 1 + 2a + pa + qb$ where h, p, q are nonnegative integers. From this fact it follows that $\gamma(u, v) = \phi + 3a - t - 1$. On the other hand, for any $x, y \in V(D)$, we have $d_{L(D)}(x, y) \leqq d(x, u_1) + d(u_1, y)$ since u_1 is both on a cycle of length a and on a cycle of length b. Now

$$d(u, u_1) = d(u, v_t) + d(v_t, u_1) = a - 1 + d(v_t, u_1) \geqq a \geqq b - 1,$$

so $d(x, u_1) \leqq d(u, u_1)$ for $x \in V(C_1)$, $d(x, u_1) \leqq a - 1 \leqq d(u, u_1)$ for $x \in V(C_2)$ and $d(x, u_1) \leqq b - 1 \leqq d(u, u_1)$ for $x \in V(C_3)$. Thus $d(x, u_1) \leqq d(u, u_1)$ for all $x \in V(D)$.

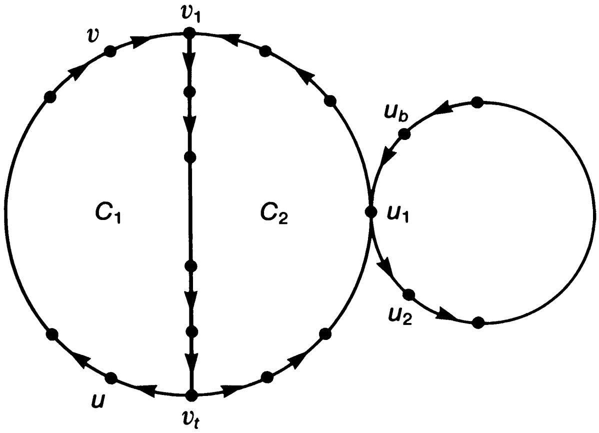


FIG. 3.2

Similarly,

$$d(u_1, v) = d(u_1, v_1) + d(v_1, v) = d(u_1, v_1) + a - 1 \geqq a \geqq b - 1$$

and it can also be checked that $d(u_1, y) \leqq d(u_1, v)$ for $y \in V(C_1)$, $V(C_2)$, $V(C_3)$, respectively. So

$$\begin{aligned} d_{L(D)}(x, y) &\leqq d(x, u_1) + d(u_1, y) \leqq d(u, u_1) + d(u_1, v) \\ &= a - 1 + d(v_t, u_1) + a - 1 + d(u_1, v_1) = 2a - 2 + d(v_t, v_1) = 3a - t - 1 \end{aligned}$$

and $\gamma(x, y) \leqq d_{L(D)}(x, y) + \phi(a, b) \leqq \phi + 3a - t - 1$ for all $x, y \in V(D)$. It follows that

$$\gamma(D) = \max_{i,j \in V(D)} \gamma(i, j) = \gamma(u, v) = \phi + 3a - t - 1 \in NE_m \subseteq NE_n.$$

This completes the proof of Lemma 3.2.

COROLLARY 3.3. *Let a, b be integers with $a \geqq b - 1$, $a, b \geqq 2$,* g.c.d$(a, b) = 1$ *and $\phi = \phi(a, b)$. If $2a + b \leqq n + 2$, then $k \in NE_n$ for all $\phi + 2a \leqq k \leqq \phi + 3a - 2$.*

Proof. Take $t = 1, 2, \cdots, a - 1$ as in Lemma 3.2.

Next we construct a new primitive, ministrong digraph and compute its exponent.

LEMMA 3.3. *Let r_1, r_2 be integers with $r_1 - 2 \geqq r_2 \geqq 2$,* g.c.d$(r_1, r_2) = 1$, *and $\phi = \phi(r_1, r_2)$. If $r_1 + r_2 + 3 \leqq n$, then $\phi + 2r_1 - 1 \in NE_n$.*

Proof. We construct a digraph D with $m = r_1 + r_2 + 3 \leqq n$ vertices consisting of a cycle $C = v_1v_2 \cdots v_{r_1+r_2}v_1$ of length $r_1 + r_2$ and three paths $v_1xv_{r_2+3}$, $v_{r_2}yv_2$, $v_{r_2-1}zv_1$ of length 2, where the vertices x, y, z are distinct and none are on the cycle C (see Fig. 3.3). Clearly D is strong. To show that D is ministrong we notice that every arc except $e = (v_1, v_2)$ either goes into a vertex of indegree 1 or comes out of a vertex of outdegree 1, so $D \backslash \{e\}$ is not strong if $e \neq (v_1, v_2)$. For $e = (v_1, v_2)$, it can be directly checked that $D \backslash \{e\}$ is also not strong, so D is ministrong. The cycle length set of D is $L(D) = \{r_1 + r_2, r_1, r_2\}$ and so D is primitive. The Frobenius number $\phi(r_1 + r_2, r_1, r_2) = \phi(r_1, r_2) = \phi$ since $r_1 + r_2$ is already a nonnegative integral linear combination of r_1 and r_2. Now we compute the exponent $\gamma(D)$. From the structure of D we see that there is a unique elementary path (of length $r_1 - 1$) from v_{r_2+1} to $v_{r_1+r_2}$ and the length of any walk from v_{r_2+1} to $v_{r_1+r_2}$ is equal to $l = r_1 - 1 + ar_1 + br_2$, where a, b are nonnegative integers and $a = 0 \Rightarrow b = 0$, so the local exponent $\gamma(v_{r_2+1}, v_{r_1+r_2}) = \phi + 2r_1 - 1$. On the other hand, for any $u, v \in L(D)$, $d_{L(D)}(u, v) \leqq d(u, v_1) + d(v_1, v)$ since the vertex v_1 is both on a cycle of length r_1 and on a cycle of length r_2. Now $d(u, v_1) \leqq r_1$ for all

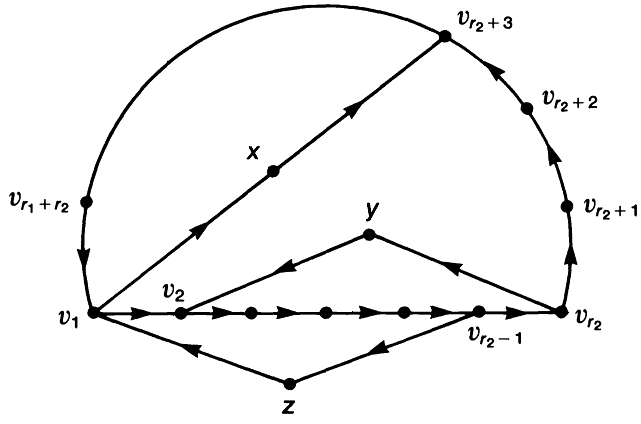


FIG. 3.3

$u \in V(D)$ (noting that $d(v_{r_2}, v_1) \leqq d(v_{r_2}, v_{r_2-1}) + 2 \leqq r_2 + 1 \leqq r_1$) and $d(v_1, v) \leqq r_1 - 1$ for all $v \in V(D)$ (noting that $d(v_1, v_{r_2+2}) = r_2 + 1 \leqq r_1 - 1$), so $d_{L(D)}(u, v) \leqq 2r_1 - 1$ and

$$\gamma(u,v) \leqq d_{L(D)}(u,v) + \phi \leqq \phi + 2r_1 - 1$$

for all u, $v \in V(D)$. It now follows that

$$\gamma(D) = \max_{i,j \in V(D)} \gamma(i,j) = \gamma(v_{r_2+1}, v_{r_1+r_2}) = \phi + 2r_1 - 1 \in NE_m \subseteq NE_n$$

and this completes the proof of Lemma 3.3.

By suitable choice of the parameters in the above lemmas and corollaries and combining the exponents, we get the following.

THEOREM 3.1. *Let r_1, r_2 be integers with $r_1 > r_2 \geqq 2$, g.c.d.$(r_1, r_2) = 1$ and $\phi = \phi(r_1, r_2)$. If $2r_1 + r_2 \leqq n$, then $k \in NE_n$ for all $\phi + 2r_1 - r_2 \leqq k \leqq \phi + 3r_1 - 2$.*

Proof. We divide the proof into the following two cases:

Case 1. $r_2 \leqq r_1 - 2$. Then $k \in NE_n$ for all $\phi + 2r_1 - r_2 \leqq k \leqq \phi + 2r_1 - 1$ by Corollary 3.1 and Lemma 3.3, and $k \in NE_n$ for all $\phi + 2r_1 \leqq k \leqq \phi + 3r_1 - 2$ by taking $a = r_1$, $b = r_2$ as in Corollary 3.3.

Case 2. $r_2 = r_1 - 1$. As in Case 1, $k \in NE_n$ for all $\phi + 2r_1 - r_2 \leqq k \leqq \phi + 2r_1 - 2$ and for all $\phi + 2r_1 \leqq k \leqq \phi + 3r_1 - 2$. To show that $\phi + 2r_1 - 1 \in NE_n$, we take $a = r_2 = r_1 - 1$, $b = r_1$ as in Corollary 3.3 to get $k \in NE_n$ for all

$$\phi + 2r_1 - 2 \leqq k \leqq \phi + 3r_1 - 5.$$

This tells us that $\phi + 2r_1 - 1 \in NE_n$ if $r_1 \geqq 4$. If $r_1 = 3$, $r_2 = 2$, then $\phi + 2r_1 - 1 = 7$. However, $n \geqq 2r_1 + r_2 = 8$, so $7 \in NE_8 \subseteq NE_n$ as well.

Combining Cases 1 and 2, we obtain Theorem 3.1.

4. The lower bound $e(n) \geqq (n^2 - 2n + 10)/9 + 1$. In this section we prove our main result. That is, for $n \geqq 8$, every integer between 6 and $(n^2 - 2n + 10)/9$ is the exponent of some primitive, ministrong digraph with n vertices.

LEMMA 4.1. *If n, x are integers with $3 \leqq x \leqq (n + 1)/3$, then $k \in NE_n$ for all $(x - 1)^2 + 2 \leqq k \leqq x^2 + 1$.*

Proof. Take $r_1 = x$, $r_2 = x - 1$ as in Theorem 3.1 to get $k \in NE_n$ for all

$$\phi(x, x-1) + x + 1 \leqq k \leqq \phi(x, x-1) + 3x - 2$$

where $\phi(x, x - 1) + x + 1 = (x - 1)^2 + 2$ and $\phi(x, x - 1) + 3x - 2 = x^2$. Take $a = x$,

$b = x - 1$ as in Corollary 3.2 to get $\phi(x, x-1) + 3x - 1 = x^2 + 1 \in NE_n$. This proves Lemma 4.1.

THEOREM 4.1. *If* $n \geqq 8$ *and* k *is any integer with* $6 \leqq k \leqq ([(n+1)/3])^2 + 1$, *then* $k \in NE_n$.

Proof. For $x = 3, 4, \cdots, [(n+1)/3]$, let $I_x = \{k \in \mathbb{Z}^+ | (x-1)^2 + 2 \leqq k \leqq x^2 + 1\}$. Then $I_x \subseteq NE_n$ by Lemma 4.1, so $I_3 \cup I_4 \cup \cdots \cup I_p \subseteq NE_n$, where $p = [(n+1)/3]$. We note that $I_{x-1} \cup I_x$ is a set of consecutive integers since $(x-1)^2 + 1$ (the upper bound of I_{x-1}) and $(x-1)^2 + 2$ (the lower bound of I_x) are consecutive integers, so $I_3 \cup I_4 \cup \cdots \cup I_p$ is also a set of consecutive integers which contains all the integers between 6 and $p^2 + 1 = ([(n+1)/3])^2 + 1$. This proves Theorem 4.1.

For the sake of simplicity, we note that

$$\left[\frac{n+1}{3}\right] \geqq \frac{n-1}{3} \quad \text{and} \quad \left(\left[\frac{n+1}{3}\right]\right)^2 + 1 \geqq \left(\frac{n-1}{3}\right)^2 + 1 = \frac{n^2 - 2n + 10}{9},$$

so that we get the lower bound $e(n) \geqq (n^2 - 2n + 10)/9 + 1$, where $e(n)$ is the least integer $\geqq 6$ such that no $n \times n$ primitive, nearly reducible matrix has exponent $e(n)$.

REFERENCES

[1] R. A. BRUALDI AND J. A. ROSS, *On the exponent of a primitive, nearly reducible matrix*, Math. Oper. Res., 5 (1980), pp. 229–241.

[2] J. A. ROSS, *On the exponent of a primitive, nearly reducible matrix*. II, this Journal 3 (1982), pp. 395–410.

[3] J. Y. SHAO, *On the exponent of a primitive digraph*, Linear Algebra Appl., 64 (1985), pp. 21–31.

[4] ———, *On a conjecture about the exponent set of primitive matrices*, Linear Algebra Appl., 65 (1985), pp. 91–123.

[5] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

006

SUBSTITUTES AND COMPLEMENTS IN CONSTRAINED LINEAR MODELS*

J. SCOTT PROVAN†

Abstract. We consider the problem of when two variables in a linear programming model can be considered to be *substitutes* (self-interfering) or *complements* (self-reinforcing). Several definitions proposed in the economic and mathematical literature are investigated in the context of linear programming models. The concept of *determinacy* is used to formalize and classify these definitions. Determinacy is studied for a class of network flow models, where graph-theoretic characterizations of substitutes and complements are given.

Key words. substitutes, complements, linear program, signsolvability

AMS(MOS) subject classifications. 05, 15, 90

1. Introduction. The study of substitutes and complements is part of the larger study of qualitative statics and complementarity in economics, and has had a long history [Go], [H], [HA], [L1], [L2], [L3], [S1], [S2]. Recently, the topic has been of interest in the mathematical literature, particularly with regard to the problem of *signsolvability* of a system of linear equations [J], [JLR], [KL], [KLM], [La], [M], [MQ], [My], [R]. The problem addressed in this paper is: When can a pair of variables be *substitutes* or *complements* in the sense of having consistently "like"/"competitive"/"self-interfering" or "opposite"/"symbiotic"/"self-reinforcing" behavior in a linear programming model? This topic has been discussed for general economic models as indicated above, and also for network models [Sh1], [Sh2], [GP], [GV] in significantly different ways. The intent of this paper is to study the concepts of substitutes and complements in the context of structural properties of a constrained linear model, and to relate these concepts to economic properties of the model. The model we consider is linear programming model

$$P(c,b):\qquad \begin{aligned} \max u(x) &= c^T x \\ Ax &= b \\ x &\geqq 0 \end{aligned}$$

where A is an $m \times n$ rank m matrix, $c \in R^n$ and $b \in R^m$. This corresponds to the economic model where a linear utility function u on n commodities is maximized over a linearly constrained commodity region. It can also be used to model the case where u is concave and piecewise linear (see, for example [Mu, § 1.2], and so covers a very rich class of economic models. We assume that $P(c, b)$ admits some optimal solution x^* with associated objective function value u^*. The study of *comparative statics* in economics involves testing the sensitivity of x^* and u^* to changes in various parameters associated with the system $P(c, b)$. Closely associated with this is the notion of *complementarity* between two variables x_i and x_j in the model, which generally speaking describes a *qualitative*, or sign consistent, relationship between the resulting changes in x_i and x_j. Samuelson [S1] provides an excellent discussion of the myriad independent and conflicting notions of complementarity given in the economic literature. All of these notions, however, seem to exhibit at least one of the following three properties: two variables x_i and x_j are called *substitutes* (*complements*, *independent*, respectively) if they:

* Received by the editors April 26, 1984; accepted for publication (in revised form) March 16, 1987. This research was supported in part by the Air Force Office of Scientific Research under contract AFOSR-84-0140.

† Department of Operations Research, University of North Carolina, Chapel Hill, North Carolina 27514.

(a) tend to replace (enhance, have no effect on) each other in the model;

(b) have like (opposite, no mutual) effect on other variables in the model;

(c) have cumulative effect on the objective function value u^* which is less than (greater than, the same as) that of the individual variables.

The purpose of this paper is to make these notions precise, and to show how they manifest themselves in the model $P(c, b)$. In the process we use the notion of *determinacy*, which was first introduced by Greenberg [Gr] and later studied in [P] and [PK]. We show how several classical definitions can be formalized in terms of determinacy, and investigate the relationships between these concepts. Finally, we consider substitutes and complements in the context of a class of network flow models, where we give an interesting graph theoretic characterization of these terms.

2. Definitions and preliminary results. Let $P(c, b)$ be as given in the Introduction. A *basis* for $P(c, b)$ is a nonsingular $m \times m$ submatrix B of A. Corresponding to each basis is the following equivalent presentation of $P(c, b)$:

$$\begin{aligned} \max u(x) = \quad & \bar{c}x_N + \bar{d} \\ (1) \qquad & x_B + \bar{A}x_N = \bar{b} \\ & x \geqq 0 \end{aligned}$$

where $x_B = (x_{B_1}, \cdots, x_{B_m})$ are the *basic* variables corresponding to the columns of B, x_N are the *nonbasic* variables corresponding to the remaining matrix N of columns of A, $\bar{c} = c_N - c_B B^{-1}N$, $\bar{d} = c_B B^{-1}b$, and $\bar{A} = B^{-1}N$. When there is no confusion, we will often denote bases by their basic variables. The *extended tableau* associated with B is the $n \times n$ matrix

$$\hat{A} = \begin{pmatrix} I & \bar{A} \\ 0 & -I \end{pmatrix}$$

whose rows and columns will be indexed by the index of the corresponding variable.

For pair x_i and x_j of variables, we call x_i and x_j *B-row substitutes* (*complements, independent*) if $\hat{a}_{ik} \cdot \hat{a}_{jk} \leqq 0$ ($\geqq 0$, $=0$) for $k = 1, \cdots, n$, and *B-column substitutes* (*complements, independent*) if $\hat{a}_{ki} \cdot \hat{a}_{kj} \geqq 0$ ($\leqq 0$, $=0$) for $k = 1, \cdots, n$. The variables x_i and x_j are *B-row* (*B-column*) *determinate* if they are B-row (B-column) substitutes, complements, or independent. When (1) is considered to determine a dependency of the basic variables x_B on changes in the nonbasic variables x_N as specified by $\bar{A}$, it becomes clear how properties (a) and (b) given in § 1 are reflected in the "row" and "column" definitions, respectively, given above. Notice by the construction of $\hat{A}$ that if $x_i = x_{B_q}$ is basic and x_j is nonbasic, then x_1 and x_j are B-row substitutes (complements, independent) if and only if they are B-column substitutes (complements, independent) if and only if $\bar{a}_{qj}$ is nonnegative (nonpositive, zero). Nonbasic variables are always B-row independent and basic variables are always B-column independent; and so the column and row definitions of substitutes and complements are consistent with respect to any basis.

Complete B-determinacy, that is, where *all* variables are either B-substitutes or B-complements, has an interesting characterization using a result of Greenberg, Lundgren and Maybee [GLM, Lemma 1]. A matrix M is *signed* if there is a subset of rows and columns of M which when negated (in any order) results in a matrix M' all of whose entries are nonnegative. The result in [GLM], when restated in the context of B-determinacy yields the following result:

PROPOSITION 2.1. *For any basis B of $P(c, b)$ and associated matrix $\bar{A}$ defined by* (1), *the following statements are equivalent*:

(i) *each pair of variables in $P(c, b)$ are B-row determinate*;

(ii) *each pair of variables in $P(c, b)$ are B-column determinate* (*in the same sense as* (i));

(iii) *$\bar{A}$ is signed.* □

We now extend the concepts of determinacy to a larger region of model activity. Let $\mathscr{B}$ be any collection of bases for $P(c, b)$. The collection $\mathscr{B}$ can be chosen to represent various types of model behavior; three important collections of bases, which will be used extensively in the paper, are given below (referring to (1)):

$\mathscr{B}_0$ = collection of *all* bases;

$\mathscr{B}_1$ = collection of *primal feasible* bases, i.e., bases for which $\bar{b} \geqq 0$;

$\mathscr{B}_2$ = collection of *dual feasible* bases, that is, bases for which $\bar{c} \leqq 0$.

Notice that the set $\mathscr{B}_1$ is independent of the choice of c and represents the set of optimal bases for $P(\gamma, b)$ as γ ranges over all values of R^n; $\mathscr{B}_2$ is independent of b and represents the set of optimal bases for $P(c, \beta)$ as β ranges over all values of R^m; and $\mathscr{B}_0$ is independent of both b and c and represents the set of optimal bases for $P(\gamma, \beta)$ as γ and β range over all values of R^n and R^m, respectively. For any basis collection, two variables x_i and x_j are called *$\mathscr{B}$-row substitutes* (etc.) if they are B-row substitutes (etc.) for all $B \in \mathscr{B}$, and *$\mathscr{B}$-row* (*column*) *determinate* if they are either B-row (column) substitutes for all $B \in \mathscr{B}$ or B-row (column) complements for all $B \in \mathscr{B}$. Note that two variables are independent—in any of the senses given above—if and only if they are both substitutes and complements; we will consequently just deal with substitutes and complements in most of the results of this paper. To differentiate from the independent case, we will sometimes call two variables *strict* substitutes (complements) if they are substitutes (complements) and not independent.

We end the section by giving two important general results. The first result concerns determinacy between collections of bases. The proof is clear.

PROPOSITION 2.2. *Let $\mathscr{B}$ and $\mathscr{B}'$ be two collections of bases for $P(c, b)$ with $\mathscr{B}'$ a subset of $\mathscr{B}$. Then any two variables which are $\mathscr{B}$-determinate are $\mathscr{B}'$-determinate. In particular, any two variables which are $\mathscr{B}_0$-determinate are $\mathscr{B}$-determinate for all collections $\mathscr{B}$ of bases.*

$\mathscr{B}_0$-determinacy was studied extensively in [P], and we make significant use of that material by using Proposition 2.2.

The next result establishes the fundamental duality between row and column determinacy. Consider the *dual* linear program to $P(c, b)$

$$D(c,b): \qquad \begin{aligned} \min v &= b^T y \\ A^T y - z &= c \\ z &\geqq 0. \end{aligned}$$

The variable z_j thus has an economic interpretation as the *negative marginal utility* (or *marginal cost*) of the variable x_j. Using the linear programming complementarity between $P(c, b)$ and $D(c, b)$, we have that the primal and dual feasible bases for $D(c, b)$ are of the form (y, z_N), where x_N is the set of nonbasic variables in some basis B for $P(c, b)$. The bases of $D(c, b)$ can be further classified by their primal or dual feasibility as follows:

$$\begin{aligned} \mathscr{B}_1^* &= \text{primal feasible bases for } D(c,b) \\ &= (y, z_N) \text{ such that } B \in \mathscr{B}_2; \\ \mathscr{B}_2^* &= \text{dual feasible bases for } D(c,b) \\ &= (y, z_N) \text{ such that } B \in \mathscr{B}_1. \end{aligned}$$

In [P, Cor. 2.6], the equivalence between $\mathscr{B}_0$-row(column) determinacy in $P(c, b)$ and $\mathscr{B}_0$-column (row) determinacy in $D(c, b)$ was established. We restate this result for substitutes and complements in the context of primal and dual feasible bases.

PROPOSITION 2.3. *Two variables z_i and z_j in $D(c, b)$ are $\mathscr{B}_1^*$-row ($\mathscr{B}_2^*$-row, $\mathscr{B}_1^*$-column, $\mathscr{B}_2^*$-column) substitutes (complements, independent) in $D(c, b)$ if and only if the corresponding variables x_i and x_j are $\mathscr{B}_2$-column ($\mathscr{B}_1$-column, $\mathscr{B}_2$-row, $\mathscr{B}_1$-row) complements (substitutes, independent) in $P(c, b)$.* □

3. Determinacy and related concepts. In this section the various forms of determinacy are put in the context of economic and geometric properties of $P(c, b)$. In order to state the results in their strongest form it will often be necessary to make a nondegeneracy assumption on $P(c, b)$ the precise form of which depends on the collection $\mathscr{B}$ of bases being considered. Referring to (1) these assumptions are:

$$\text{(N1)} \qquad \bar{b} > 0 \quad \text{for all } B \in \mathscr{B}_1,$$

$$\text{(N2)} \qquad \bar{c} < 0 \quad \text{for all } B \in \mathscr{B}_2.$$

Note that $P(c, b)$ can always be made to satisfy (N1) or (N2) by perturbing slightly b or c, respectively. The first result of the section concerns the qualitative structure of the feasible region for $P(c, b)$, and the sensitivity of the optimal solution under parameterization of costs. It also relates to work of Granot and Veinott [GV] which will be discussed in more detail later. We assume that the feasible region of $P(c, b)$ is bounded; although the concepts given here can be extended to the unbounded case, they become so cumbersome as to lose their intuitive appeal. An *edge* of $P(c, b)$ is defined to be an edge of the feasible region taken as a polytope (see [Mu, § 3.6]). Note that edges have affine dimension one, so that relative changes between variables on an edge can be uniquely determined. For $\gamma \in R^n$, define the two functions $x^*(\gamma)$ and $z^*(\gamma)$ as follows:

$$x^*(\gamma) = (x_1^*(\gamma), \cdots, x_n^*(\gamma)) = \text{the optimal solution to the linear program } P(\gamma, b),$$

$$u^*(\gamma) = \gamma \cdot x^*(\gamma) = \text{the optimal objective function value for } P(\gamma, b).$$

Let Ω^* be the set of values of γ for which $x^*(\gamma)$ is uniquely defined so that Ω^* is an open set whose complement is of measure zero. Note that $u^*(\gamma)$ is defined for all $\gamma \in R^n$, and is continuous. The function $x_i^*(\gamma)$ is said to be *nonincreasing* (*nondecreasing*) in γ_j if for every γ^1, $\gamma^2 \in \Omega^*$ with $\gamma^1 \leqq \gamma^2$ and γ^1 and γ^2 differing only on the jth component we have

$$x_i^*(\gamma^2) \leqq (\geqq) x_i^*(\gamma^1).$$

The function $u^*(\gamma)$ is said to be *submodular* (*supermodular*) in γ_i and γ_j if for every γ^1, $\gamma^2 \in R^n$ with $\gamma^1 \leqq \gamma^2$ and γ^1 and γ^2 differing only in the ith and jth components, we have

$$u^*(\gamma^1) + u^*(\gamma^2) \leqq (\geqq) u^*(\cdots, \gamma_i^1, \cdots, \gamma_j^2, \cdots) + u^*(\cdots, \gamma_i^2, \cdots, \gamma_j^1, \cdots).$$

Sub(super)modularity of u^* corresponds to property (c) of substitutes (complements) given in the first section. It says in essence that the cumulative effect on u^* of increasing both the ith and jth component is less than or equal to (greater than or equal to) the sum effects of increasing each component individually, or equivalently, that an increase in either of the components can only decrease (increase) the marginal effect on u^* of increasing the other component. Thus it is the analogue in a nondifferentiable setting of the second partial derivative of u^* with respect to x_i and x_j being everywhere nonpositive (nonnegative). We now have the following result:

THEOREM 3.1. *Suppose $P(c, b)$ satisfies nondegeneracy condition* (N1) *and has a bounded feasible region. Then for variables x_i and x_j in $P(c, b)$, the following statements are equivalent*:

(i) *x_i and x_j are $\mathscr{B}_1$-row substitutes* (*complements*);

(ii) *$\Delta x_i/\Delta x_j$ is nonpositive* (*nonnegative*) *over all edges of $P(c, b)$ for which x_j is not constant*;

(iii) *$x_i^*(\gamma)$ is nonincreasing* (*nondecreasing*) *in γ_j*;

(iv) *$u^*(\gamma)$ is submodular* (*supermodular*) *in γ_i and γ_j*.

Proof. We will prove all equivalences for the case of substitutes, the case for complements being symmetric.

(i) ⇒ (ii): From the nondegeneracy assumptions on $P(c, b)$ and standard results in linear programming theory ([Mu, § 3.6]) it follows that each edge of $P(c, b)$ can be described by giving a basis $B \in \mathscr{B}_1$ and column k, and then defining the edge resulting from a simplex pivot on the tableau defined by (1) in column k. The resulting edge is of the form

$$e = \{x^0 - \lambda \hat{A}^k : 0 \leqq \lambda \leqq \bar{b}_r/\bar{a}_{rk}\}$$

where x_0 is the basic solution corresponding to basis B, $\hat{A}^k$ is the kth column of $\hat{A}$ and the pivot occurred on element $\bar{a}_{rk}$. It follows that if x_j is not constant on e then $\Delta x_i/\Delta x_j = \hat{a}_{ik}/\hat{a}_{jk}$, which is nonpositive if x_i and x_j are $\mathscr{B}_1$-row substitutes.

(ii) ⇒ (iii): Let $\gamma^0, \gamma^t \in \Omega^*$ with $\gamma^0 \leqq \gamma^t$, γ^0 and γ^t differing only on the jth coordinate, and let $x^0 = x^*(\gamma^0)$, $x^t = x^*(\gamma^t)$. By perturbing γ^0 and γ^t slightly, we can assume that a path of basic feasible solutions $x^0, x^1, \cdots, x^r = x^t$ exists satisfying $x^k = x^*(\gamma^k)$, $k = 0, \cdots, r$, with $\gamma^k = (\gamma_1^0, \cdots, \gamma_j^0 + \alpha_k, \cdots, \gamma_n^0) \in \Omega^*$, $0 = \alpha_1 < \cdots < \alpha_r = \gamma_j^t - \gamma_j^0$ and such that consecutive x^k are joined by an edge of $P(c, b)$. Consider one such edge (x^k, x^{k+1}). Since γ_j is increasing between γ^k and γ^{k+1}, it must be that x_j is also increasing between γ^k and γ^{k+1}. It follows that if $\Delta x_i/\Delta x_j$ is nonpositive along the associated edge, then $x_i^{k+1} \leqq x_i^k$. Thus $x_i^*(\gamma^0) = x_i^0 \leqq x_i^t = x_i^*(\gamma^t)$ and (iii) follows.

(iii) ⇒ (iv): Let $\gamma^1, \gamma^2 \in R^n$ with $\gamma^1 \leqq \gamma^2$ and γ^1 and γ^2 differing only on the ith and jth coordinates. For $\gamma_i^1 = \alpha_0 \leqq \alpha \leqq \alpha_t = \gamma_i^2$, define the two n-vectors $\bar{\gamma}^1(\alpha)$ and $\bar{\gamma}^2(\alpha)$ such that for $p = 1, 2$, $\bar{\gamma}^p(\alpha)$ agrees with γ^p except on the ith coordinate, where it has the value α. Then $\bar{\gamma}^1(\alpha_0) = \gamma^1$, $\bar{\gamma}^1(\alpha_t) = (\cdots, \gamma_i^2, \cdots, \gamma_j^1, \cdots)$, $\bar{\gamma}^2(\alpha_0) = (\cdots, \gamma_i^1, \cdots, \gamma_j^2, \cdots)$, and $\bar{\gamma}^2(\alpha_t) = \gamma^2$. Further, for all $\alpha_0 \leqq \alpha \leqq \alpha_t$, $\bar{\gamma}^1(\alpha) \leqq \bar{\gamma}^2(\alpha)$ and $\bar{\gamma}^1(\alpha)$ and $\bar{\gamma}^2(\alpha)$ differ only on the jth coordinate. Let x^{pq} be an optimal solution to $P(\bar{\gamma}^p(\alpha_q), b)$ for $p = 0, 1$ and $q = 0, t$. By varying α parametrically from α_0 to α_t we obtain two paths of optimal points—one from x^{10} to x^{1t} via $\bar{\gamma}^1(\alpha)$ and the other from x^{20} to x^{2t} via $\bar{\gamma}^2(\alpha)$—with pivots occurring in one or the other of these paths only at specified parameter values $\alpha_1, \cdots, \alpha_r$ with $\alpha_0 \leqq \alpha_1 < \cdots \alpha_{r-1} \leqq \alpha_r = \alpha_t$. Thus for $k = 1, \cdots, r$, $x^*(\bar{\gamma}^1(\alpha))$ and $x^*(\bar{\gamma}^2(\alpha))$ are both constant over the range $\alpha_{k-1} < \alpha < \alpha_k$. By perturbing the objective functions slightly (and appealing to the continuity of u^*) we can assume that $\bar{\gamma}^1(\alpha)$ and $\bar{\gamma}^2(\alpha)$ are both in Ω^* in the range $\alpha_{k-1} < \alpha < \alpha_k$. Therefore (iii) applies and so $x_i^*(\bar{\gamma}^1(\alpha)) \geqq x_i^*(\bar{\gamma}^2(\alpha))$. But now for any pair $\alpha_{k-1} < \alpha < \beta < \alpha_k$ we have

$$\begin{aligned} u^*(\bar{\gamma}^2(\beta)) - u^*(\bar{\gamma}^2(\alpha)) &= (\beta - \alpha) x_i^*(\bar{\gamma}^2(\alpha)) \\ &\leqq (\beta - \alpha) x_i^*(\bar{\gamma}^1(\alpha)) = u^*(\bar{\gamma}^1(\beta)) - u^*(\bar{\gamma}^1(\alpha)) \end{aligned}$$

and so by the continuity of u^*,

$$u^*(\bar{\gamma}^2(\alpha_k)) - u^*(\bar{\gamma}^2(\alpha_{k-1})) \leqq u^*(\bar{\gamma}^1(\alpha_k)) - u^*(\bar{\gamma}^1(\alpha_{k-1})).$$

By summing these inequalities over $k = 1, \cdots, r$ we get

$$\begin{aligned} u^*(\gamma^2) - u^*(\cdots, \gamma_i^2, \cdots, \gamma_j^1, \cdots) &= u^*(\bar{\gamma}^2(\alpha_t)) - u^*(\bar{\gamma}^2(\alpha_0)) \\ &\leqq u^*(\bar{\gamma}^1(\alpha_t)) - u^*(\bar{\gamma}^1(\alpha_0)) \\ &= u^*(\cdots, \gamma_i^1, \cdots, \gamma_j^2, \cdots) - u^*(\gamma^1) \end{aligned}$$

and (iv) follows.

(iv) ⇒ (i): Let $B \in \mathscr{B}_1$ with $\hat{A}$ the associated tableau, and let i and j be two rows of $\hat{A}$. If x_i and x_j are both nonbasic, then $\hat{a}_{ik} \cdot \hat{a}_{jk} = 0$ for all k. Otherwise assume by symmetry that $x_j = x_{B_q}$ is basic, and let k be any column index. If either $\hat{a}_{ik}$ or $\hat{a}_{jk}$ equals zero then $\hat{a}_{ik} \cdot \hat{a}_{jk} = 0$. Otherwise we have that k is a nonbasic column, with $\hat{a}_{jk} = \bar{a}_{qk} \neq 0 \neq \hat{a}_{ik}$. We take first the case where $\bar{a}_{qk} > 0$. Define the n-vector γ^{11} by

$$\gamma_p^{11} = \begin{cases} \bar{a}_{qp} - 1, & p \in N - \{k\}, \\ \bar{a}_{qk} - \varepsilon/2, & p = k, \\ 1, & p = j, \\ 0 & \text{otherwise} \end{cases}$$

with ε a small positive scalar. Then

$$\gamma_N^{11} - \gamma_B^{11}\bar{A} = (-1, \cdots, \overset{k}{-\varepsilon/2}, \cdots, -1)$$

and so $\gamma^{11} \in \Omega^*$ with B the basis corresponding to $x^*(\gamma^{11})$. Define γ^{12} by

$$\gamma_p^{12} = \begin{cases} \gamma_j^{11} + \varepsilon/\bar{a}_{qk}, & p = j, \\ \gamma_p^{11} & \text{otherwise} \end{cases}$$

so that

$$\gamma_N^{12} - \gamma_B^{12}\bar{A} = (-1 + \varepsilon\bar{a}_{q1}/\bar{a}_{qk}, \cdots, \varepsilon/2, \cdots, -1 + \varepsilon\bar{a}_{qn}/\bar{a}_{qk}).$$

With ε small enough, this plus the boundedness of $P(c, b)$ insures $\gamma^{12} \in \Omega^*$ with optimal basis B' for $P(\gamma^{12}, b)$ satisfying $B' = B' \cup \{k\} - \{l\}$ for some basic index $l = B_s$. Now choose $\delta > 0$ sufficiently small so that for $r = 1, 2$ the n-vector γ^{2r} defined

$$\gamma_p^{2r} = \begin{cases} \gamma_i^{1r} + \delta, & p = i, \\ \gamma_p^{1r} & \text{otherwise} \end{cases}$$

is also in Ω^* with

$$x^*(\gamma^{2r}) = x^*(\gamma^{1r}).$$

Then $\gamma^{11} \leqq \gamma^{22}$ and γ^{11} and γ^{22} differ only on coordinates i and j. Thus by (iv) we have

$$u^*(\gamma^{11}) + u^*(\gamma^{22}) \leqq u^*(\gamma^{21}) + u^*(\gamma^{12}),$$

that is,

$$\begin{aligned} u^*(\gamma^{22}) - u^*(\gamma^{12}) &= \delta x_i^*(\gamma^{12}) \\ &\leqq u^*(\gamma^{21}) - u^*(\gamma^{11}) = \delta x_i^*(\gamma^{11}) \end{aligned}$$

implying $x_i^*(\gamma^{12}) - x_i^*(\gamma^{11}) \leqq 0$.

But

$$x_i^*(\gamma^{12}) - x_i^*(\gamma^{11}) = -(\bar{b}_s/\bar{a}_{sk})\hat{a}_{ik}$$

where the pivot occurred on element $\bar{a}_{sk}$. The nondegeneracy assumption insures that $\bar{b}_s/\bar{a}_{sk} > 0$, and so $\hat{a}_{ik} \geqq 0$ and thus $\hat{a}_{ik} \cdot \hat{a}_{jk} \geqq 0$. The case when $\bar{a}_{qk} < 0$ is handled similarly, choosing ε and δ less than zero and swapping the roles of γ^{11} and γ^{22}. By doing this for all $B \in \mathscr{B}_1$, and all columns k it follows that x_i and x_j are $\mathscr{B}_1$-row substitutes. This completes the proof of the theorem. □

A slightly weaker version of Theorem 3.1 applies when we drop the boundedness and degeneracy conditions on $P(c, b)$. Then the edges of $P(c, b)$ include extreme rays, and the functions $x^*(\gamma)$ and $u^*(\gamma)$ are defined only over γ for which $P(\gamma, b)$ is a bounded linear program. The following corollary is now straightforward.

COROLLARY 3.2. *Without the conditions on* $P(c, b)$ *given in Theorem* 3.1, *we still have* (i) ⇒ (ii) ⇒ (iii) ⇒ (iv). □

The next result of this section concerns the sensitivity of the optimal solution of $P(c, b)$ under parameterization of the variables, and is a direct generalization of work by Gale and Politif [GP] and Shapley [Sh1], [Sh2]. Here we presume that $P(c, b)$ satisfies nondegeneracy assumption (N2), and that $P(c, b)$ is feasible for all $\beta \in R^m$ (or equivalently, that the dual program $D(c, b)$ has a bounded feasible region). For $\alpha \in R^n$ consider the linear program obtained by parameterizing variable values through setting lower bounds α_j on x_j, $j = 1, \cdots, n$:

$$\begin{array}{ll} & \max u = cx \\ P'(\alpha): & \quad Ax = b \\ & \quad x \geqq \alpha. \end{array}$$

The assumptions on $P(c, b)$ insures that $P'(\alpha)$ has a unique optimal solution for all α, which we shall denote $x^{**}(\alpha)$, with associated objective function denoted $u^{**}(\alpha)$. We can now state a similar result to that of Theorem 3.1 for x^{**} and u^{**}. For variable x_i in $P(c, b)$ denote by $\mathscr{B}_2^i$ the set of all bases in $\mathscr{B}_2$ for which x_i is basic.

THEOREM 3.3. *Suppose* $P(c, b)$ *satisfies nondegeneracy condition* (N2) *and the property that* $P(c, \beta)$ *is feasible for all* $\beta \in R^m$. *Then*

(i) $x_i^{**}(\alpha)$ *is nonincreasing* (*nondecreasing*) *in* α_j *if and only if* x_i *and* x_j *are* $\mathscr{B}_2^i$-*column substitutes* (*complements*);

(ii) $u^{**}(\alpha)$ *is submodular* (*supermodular*) *in* α_i *and* α_j *if and only if* x_i *and* x_j *are* $\mathscr{B}_2$-*column substitutes* (*complements*).

Further, (ii) *implies* (i).

Proof. Again, we prove the equivalences for the case of substitutes, the case for complements being symmetric.

(i) (⇐): Let $\alpha^0, \alpha^t \in R^n$ be given with $\alpha^0 \leqq \alpha^t$ and α^0 and α^t differing only on the jth coordinate. For $0 \leqq \gamma \leqq \alpha_j^t - \alpha_j^0$ define $\bar{\alpha}(\gamma) = (\alpha_1^0, \cdots, \alpha_j^0 + \gamma, \cdots, \alpha_n^0)$, so that $\bar{\alpha}(0) = \alpha^0$ and $\bar{\alpha}(\alpha_j^t - \alpha_j^0) = \alpha^t$. By varying γ from 0 to $\alpha_j^t - \alpha_j^0$ we obtain a sequence $0 = \gamma_0 \leqq \gamma_1 \leqq \cdots \leqq \gamma_r = \alpha_j^t - \alpha_j^0$ such that $x^{**}(\bar{\alpha}(\gamma))$ has the same basis $B^p \in \mathscr{B}_2$ for $\gamma_p < \gamma < \gamma_{p+1}$, and dual simplex pivots occur at each γ_i, $i = 1, \cdots, r-1$. We can therefore concentrate on the change in $x_i^{**}(\bar{\alpha}(\gamma))$ for $\gamma_p \leqq \gamma \leqq \gamma_{p+1}$. If either x_i is nonbasic or x_j is basic with respect to B^p, then $x_i^{**}(\bar{\alpha}(\gamma))$ is constant in the interval $\gamma_p \leqq \gamma \leqq \gamma_{p+1}$ and we are done. Otherwise we have $\Delta x^{**}(\bar{\alpha}(\gamma))/\Delta\gamma = -\bar{a}_{qj}$, where $i = B_q^p$. Therefore, if x_i and x_j are B^p-column substitutes then $-\bar{a}_{qj} \leqq 0$, and so $x_i^{**}(\bar{\alpha}(\gamma))$ is nonincreasing in the interval $\gamma_p \leqq \gamma \leqq \gamma_{p+1}$. By performing this over each interval, we get that $x_i^{**}(\alpha^0) \leqq x_i^{**}(\alpha^t)$ and the implication follows.

(i) (⇒): Let B be an element of $\mathscr{B}_2^i$. Then x_i is basic in B. If x_j is also basic in B, then x_i and x_j are B-column independent. Otherwise define $\alpha = (\alpha_B, \alpha_N)$ with $\alpha_N = 0$

and $\alpha_B = B^{-1}b - e$, where e is the vector of ones. Then $P'(\alpha)$ can be written

$$\begin{aligned} &\max c^T x' + c^T \alpha \\ &Ax' = Be \\ &x' \geqq 0 \end{aligned}$$

where $x' = x - \alpha$. The basic feasible solution (x'_B, x'_N) for this program associated with B has $x'_B = e > 0$, and since $B \in \mathscr{B}_2$ then B is the optimal basis for $P'(\alpha)$. Now let α' be obtained from α by increasing the jth coordinate by a small amount ε. Then $x_i^{**}(\alpha') = x_i^{**}(\alpha) - \bar{a}_{qj}\varepsilon$, where $i = B_q$. Since $x_i^{**}(\alpha)$ is nonincreasing in α_j we must have $\bar{a}_{qj} \geqq 0$, and thus x_i and x_j are B-column substitutes. This completes the proof of (i).

(ii): Here we can simply look at the dual program, and apply Proposition 2.3 and Theorem 3.1. The dual program to $P'(\alpha)$ (in maximization form) is

$$D'(\alpha): \quad \begin{aligned} \max v = -b^T y + \alpha^T z& \\ A^T y - \quad z = c& \\ z \geqq 0& \end{aligned}$$

with the optimal value $v^*(\alpha)$ equal to $-u^{**}(\alpha)$. The submodularity of $u^{**}(\alpha)$ is then equivalent to the *super*modularity of $v^*(\alpha)$. Now the conditions on $P(c, b)$ in the theorem are precisely those necessary for the conditions of Theorem 3.1 to hold for $D'(\alpha)$. Thus we have $v^*(\alpha)$ supermodular in α_i and α_j if and only if z_i and z_j are $\mathscr{B}_1^*$-row complements. But by Proposition 2.3 this is true if and only if x_i and x_j are $\mathscr{B}_2$-column substitutes, and the equivalence follows.

That (ii) implies (i) follows immediately from Proposition 2.2 and the fact that $\mathscr{B}_2^i$ is a subset of $\mathscr{B}_2$, and this completes the proof of the theorem. □

If the restrictions on $P(c, b)$ in Theorem 3.3 are dropped, we still obtain a weakened version, by restricting the domain of x^{**} and u^{**} to values of α for which $P'(\alpha)$ is feasible. Similar to Corollary 3.2 we have

COROLLARY 3.4. *Without the restrictions on* $P(c, b)$ *given in Theorem* 3.3, *the "if" portions of* (i) *and* (ii) *still hold.* □

Several comments are in order. First, if nondegeneracy condition (N1) is also present for $P(c, b)$, then Theorem 3.3 continues to apply when the domains of x^{**} and u^{**} are restricted to nonnegative values of α. Thus we may treat the inequalities $x \geqq \alpha$ as a restriction of the feasible region of $P(c, b)$ (see Theorem 3.5). Second, note that statement (ii) in Theorem 3.3 strictly implies statement (i). This can be seen by considering the following linear program:

$$\begin{array}{rllllll} \max u = & & & x_3 & -x_4 & -x_5 & -3x_6 \\ & x_1 & & -x_3 & & +x_5 & +x_6 = 1 \\ & & x_2 & & -x_4 & +x_5 & -x_6 = 1 \\ & & & & x_1, \cdots, x_6 \geqq 0. & & \end{array}$$

This program satisfies the conditions of Theorem 3.3. Further, x_5 and x_6 are $\mathscr{B}_2^5$- and $\mathscr{B}_2^6$-column independent since they are nonbasic in every basis of $\mathscr{B}_2$, and so $x_5^{**}(\alpha)$ is constant in α_6 and vice versa. But x_5 and x_6 are not $\mathscr{B}_2$-determinate (take basis (x_1, x_2)) and so $u^{**}(\alpha)$ is neither supermodular nor submodular in α_5 and α_6. As a final note, Example 4.8 in the final section will show that the conditions of Theorems 3.1 and 3.3 are not equivalent, that is, $\mathscr{B}_1$-row substitutes (complements) are not necessarily $\mathscr{B}_2$-column substitutes (complements).

We end the section by showing how the concepts of determinacy relate to the results of Granot and Veinott [GV, Thm. 17 and Cor. 19] on substitutes and complements. The results here are simplified somewhat to avoid added notation, although most of the more general results in that paper also apply in this context. Let $P(c, b)$ have a bounded feasible region. Consider the following optimization problem:

$$P^f(t): \qquad \begin{aligned} \max u = \sum_{j=1}^{n} f_j(x_j, t_j) \\ Ax = b \\ x \geqq 0 \end{aligned}$$

where each $f_j: R^2 \to R \cup \{-\infty\}$ is concave and lower semicontinuous in t_j and supermodular in x_j and t_j (with the obvious extensions when $f(x_j, t_j) = -\infty$). As was done for Theorem 3.1, define $x^f(t)$ to be the optimal solution, and $u^f(t)$ the optimal objective function value, for $P^f(t)$. Note that u^f is defined for all $t \in R^n$ and $x^f(t)$ is defined over subset Ω^f of R^n. It turns out that the analogous result for Theorem 3.1 involves $\mathscr{B}_0$-determinacy. From Proposition 2.2 it follows that $\mathscr{B}_0$-determinacy implies $\mathscr{B}_1$- and $\mathscr{B}_2$-determinacy, and so the result which follows will imply the conditions of Theorems 3.1 and 3.3 and their corollaries. Further, in [P, Thm. 2.12] it was proved that $\mathscr{B}_0$-row determinacy and $\mathscr{B}_0$-column determinacy are equivalent (in the same sense). Thus we will drop the "row" and "column" labels in the following discussion.

THEOREM 3.5. *Let $P(c, b)$ have a bounded feasible region with at least one strictly positive solution, and let x_i and x_j be two variables in $P(c, b)$. Then the following are equivalent*:

(i) *x_i and x_j are $\mathscr{B}_0$-substitutes (complements)*;

(ii) *$x_i^f(t)$ is nonincreasing (nondecreasing) in t_j for every $f_1, \cdots, f_n$ as defined for $P^f(t)$*;

(iii) *$u^f(t)$ is submodular (supermodular) in t_i and t_j for every $f_1, \cdots, f_n$ as defined for $P^f(t)$.*

Proof. The proofs (i) ⇒ (ii) and (i) ⇒ (iii) are straightforward extensions of those of Theorem 10, Theorem 17 and Corollary 19 in [GV] (with the terms "submodular" ("supermodular") in place of "superadditive" ("subadditive") since that paper considers the minimization problem). The key property of x_i and x_j needed for those proofs is that for each element y of the null space of A whose support is minimal, $y_i \cdot y_j$ is *nonpositive* in the case of substitutes and *nonnegative* in the case of complements. Let C be the support of such a y. If y_j is not in C, then $y_i \cdot y_j = 0$. Otherwise let A_C be the corresponding set of columns of A. Since y has minimal support then removing the jth column from A_C results in an independent set of columns which can be extended to a basis $B \in \mathscr{B}_0$ not containing the jth column. But now if we consider the tableau $\bar{A}$ associated with B with $q = B_i$ then we have $y_i = -\bar{a}_{qj} \cdot y_j$ and so $y_i \cdot y_j = -\bar{a}_{qj}$ is nonpositive if x_i and x_j are B-substitutes and nonnegative if x_i and x_j are B-complements. The remainder of the proof is exactly as that of the theorems in [GV].

For the proofs (ii) ⇒ (i) and (iii) ⇒ (ii), choose $B \in \mathscr{B}_0$. Let β be any element of R^m and let $x^0 > 0$ be feasible to $P(c, b)$ as specified by the theorem. Define $\alpha = (\alpha_B, \alpha_N)$ with $\alpha_N = x_N^0$ and $\alpha_B = x_B^0 - \varepsilon B^{-1}\beta$, where $\varepsilon > 0$ is chosen small enough so that α_B is nonnegative. Then the system

$$Ax = b,$$
$$x \geqq \alpha$$

is equivalent to the system

$$Ax' = \beta,$$
$$x' \geqq 0$$

with $x' = (x - \alpha)/\varepsilon$. Now for any $\gamma \in R^n$ define the functions f_j, $j = 1, \cdots, n$, by

$$f_j(x_j, t_j) = \gamma_j x_j + \delta_-(x_j - \alpha_j) + t_j$$

where $\delta_-(y) = -\infty$ if $y \leqq 0$ and 0 otherwise. Then $f_1, \cdots, f_n$ satisfy the conditions for $P^f(t)$, and the optimal solution for $P^f(t)$ will be equivalent to that for $P(\gamma + t, \beta)$ up to translation by α and positive scalar multiple $1/\varepsilon$. What we have, therefore, is the situation in Theorem 3.1, with (iii) and (iv) of that theorem implied by (ii) and (iii), respectively, of the present theorem. It follows that x_i and x_j are $\mathscr{B}_1$-row substitutes (complements) for $\mathscr{B}_1$ associated with $P(c, \beta)$, and so by ranging β over all vectors in R^m we have that x_i and x_j are $\mathscr{B}_0$-substitutes (complements). This proves the theorem.

4. Determinacy in a class of network models. Complementarity in network models has been the subject of considerable study [GP], [GV], [Sh1], [Sh2]. Here we consider $\mathscr{B}_1$-row determinacy—and hence any of the equivalent properties of Theorem 3.1—in the context of a particular class of network models, and give graph-theoretic characterizations of the associated properties. Define a *transshipment matrix* to be an $m \times n$ $(0, \pm 1)$ matrix A with exactly one $+1$, exactly one -1, or exactly one $+ 1$ and one -1 in each column; and a *transshipment* model as any system $P(c, b)$ with A a transshipment matrix. Transshipment models occur in numerous network related problems, most notably transportation and network flow problems. Associated with any transshipment matrix A is a directed network $G(A) = (V, E)$ whose node set V corresponds to the rows of A together with an additional *source node* r, and whose arc set E corresponds to columns of A, where, for $k = 1, \cdots, n$, the arc associated with x_k is

$$e_k = \begin{cases} (v_i, v_j) & \text{if } a_{ik} = -1 \text{ and } a_{jk} = +1, \\ (r, v_j) & \text{if } a_{jk} = +1 \text{ and } a_{lk} = 0 \text{ for } l \neq j, \\ (v_i, r) & \text{if } a_{ik} = -1 \text{ and } a_{lk} = 0 \text{ for } l \neq i. \end{cases}$$

A transshipment matrix A has two interesting properties which relate to the previous two sections:

(1) Leontief property—each column of A contains at most one positive element;

(2) Totally unimodular property—every square submatrix of A has determinant $+1$, -1, or 0.

The Leontief property allows us to consider simultaneously a large class of models $P(c, b)$ associated with the matrix A. This can be done by using the following standard result in Leontief theory (see [KWW, Thm. 2.3.4].

LEMMA 4.1. *For any $m \times n$ rank m Leontief matrix A and any set B of columns of A, the following are equivalent:*

(i) *for some fixed positive b, there is an $x_B \geqq 0$ such that $b = Bx_B$;*

(ii) *for every positive b there is an $x_B \geqq 0$ such that $b = Bx_B$.* □

From Lemma 4.1 we have the following result, whose proof is immediate.

THEOREM 4.2. *Let A be an $m \times n$ rank m Leontief matrix, and $P(c, b)$ the associated linear program where b is a positive vector. Then $P(c, b)$ will always satisfy nondegeneracy assumption* (N1), *and the basis collection $\mathscr{B}_1$ is independent of the particular values of c or (positive) b.* □

The remainder of the section will be spent characterizing $\mathscr{B}_1$-row determinacy for any—and hence all—of the transshipment models $P(c, b)$ with $b > 0$. We will refer to these as *reachability models*. The collection $\mathscr{B}_1$ in a reachability model has a useful characterization in terms of properties of the graph $G(A)$. Define a *spanning tree* in a graph G to be set T of arcs of G for which there is a unique (undirected) path $\Gamma(T, v)$ from the root node r to every node v of G. The tree T is called an *r-rooted spanning arborescence*, or simply *r-tree*, if the path $\Gamma(T, v)$ is always directed from r to v. Equivalently, T is an r-tree if T is a spanning tree and there is at least one arc, and hence exactly one arc, of T pointing into every node of $V\backslash\{r\}$. For any subset S of nodes containing r, we define an r-tree *on S* to be an r-tree on the subnetwork generated by S. The next result follows from standard network theory.

THEOREM 4.3. *Let $P(c, b)$ be a reachability model and B a set of columns of the associated matrix A. Then B is in $\mathscr{B}_1$ if and only if the arc set T_B corresponding to B forms an r-tree in $G(A)$.* □

From Theorem 4.3 it follows that the collection $\mathscr{B}_1$ of a reachability model is nonempty if and only if G is r-*connected*, that is, there exists a directed path from r to every node of G. Henceforth, we will consider only r-connected graphs. The second property of A, that of being totally unimodular, provides a useful property to have when studying activity with respect to a particular basis B. Its proof is in [P, Prop. 3.1].

PROPOSITION 4.4. *Let $P(c, b)$ have A be totally unimodular, and let B be any basis for $P(c, b)$. Then every two variables are either B-substitutes or B-complements.* □

From this we have the following corollary.

COROLLARY 4.5. *For any system $P(c, b)$ with A totally unimodular and any collection $\mathscr{B}$ of bases, two variables x_i and x_j are B-substitutes (B-complements) in either sense if and only if there is no basis B in $\mathscr{B}$ for which x_i and x_j are strict B-complements (B-substitutes). They are independent if and only if there is no B in $\mathscr{B}$ for which x_i and x_j are either strict B-substitutes or strict B-complements.* □

We now consider a particular basis B, with corresponding r-tree T_B in $G(A)$, and identify the values of the matrix $\bar{A}$. This was done in [P, Lemma 3.4] for the basis collection $\mathscr{B}_0$, but we restate it here in the special case when B is in $\mathscr{B}_1$. To find the value $\bar{a}_{qk}$ with $i = B_q$, we add the corresponding edge $e_k = (u, v)$ to the r-tree T_B, forming a unique *fundamental circuit* $C(T_B, e_k)$. This circuit is in turn partitioned into two *parts*, namely,

$$C_1(T_B, e_k) = \Gamma(T_B, v) \cap C(T_B, e_k),$$

$$C_2(T_B, e_k) = \Gamma(T_B, u) \cap C(T_B, e_k) \cup \{e_k\}.$$

From Lemma 3.4 in [P], we obtain the value of $\bar{a}_{qk}$ as

$$\bar{a}_{qk} = \begin{cases} +1 & \text{if } e_i \in C_1(T_B, e_k), \\ -1 & \text{if } e_i \in C_2(T_B, e_k), \\ 0 & \text{if } e_i \notin C(T_B, e_k). \end{cases}$$

This results in the following lemma.

LEMMA 4.6. *Let $P(c, b)$ be a reachability model, B a basis in $\mathscr{B}_1$, and i and j distinct indices.*

(i) *If i and j are both nonbasic then x_i and x_j are B-row independent;*

(ii) *if i is basic and j is nonbasic, then x_i and x_j are strict B-row substitutes if and only if $e_i \in C_1(T_B, e_j)$ and strict B-row complements if and only if $e_i \in C_2(T_B, e_j)$;*

(iii) *if i and j are both basic, then x_i and x_j are strict B-row substitutes if and only if there is an arc e_k such that e_i and e_j are different parts of $C(T_B, e_k)$, and strict B-row complements if and only if there is an arc e_k such that e_i and e_j are on the same part of $C(T_B, e_k)$.* □

Corollary 4.5 and Lemma 4.6 allow us to obtain a graph-theoretic characterization for row substitutes and complements in the model $P(c, b)$. Define an *r-lasso* $L = \Gamma \cup C_1 \cup C_2$ to consist of directed paths Γ, C_1, and C_2, node disjoint except for nodes u and t, such that Γ goes from r to u and C_1 and C_2 each go from u to t (see Fig. 1). We allow u to equal r, in which case Γ is empty, and t to equal u, in which case one of C_1 and C_2 is empty. The node t is called the *endpoint* of the lasso, and the final arcs on C_1 and C_2 are called *end arcs* of the lasso. It follows that there is a unique path from r to any node of L except the endpoint, and that removal of either end arc of L creates an r-tree on the nodes of L. Two arcs e_i and e_j are said to be on the *same* (*opposite*) side of L if they lie on the same (opposite) C_i. We can now give the characterization for $\mathscr{B}_1$-row substitutes and complements.

THEOREM 4.7. *Let $P(c, b)$ be a reachability system, and let x_i and x_j be two variables. Then x_i and x_j are $\mathscr{B}_1$-row substitutes* (*complements*) *if and only if there is no r-lasso in $G(A)$ with e_i and e_j on the same* (*opposite*) *sides.*

Proof. We prove the theorem for substitutes, the argument for complements being symmetric. By Corollary 4.5, x_i and x_j are $\mathscr{B}_1$-row substitutes if and only if there is no basis $B \in \mathscr{B}_1$ for which x_i and x_j are strict B-row complements. Suppose such a basis exists, and apply Lemma 4.6. Then at least one of x_i and x_j must be basic. If exactly one of i and j are basic, say i is basic and j is nonbasic, then x_i and x_j are strict B-row complements if and only if $e_i \in C_2(T_B, e_j)$. Letting $e_j = (u, v)$, we have that the set $L = \Gamma(T_B, u) \cup \Gamma(T_B, v) \cup \{e_j\}$ is an r-lasso in $G(A)$, and e_i and e_j are on the same side of L.

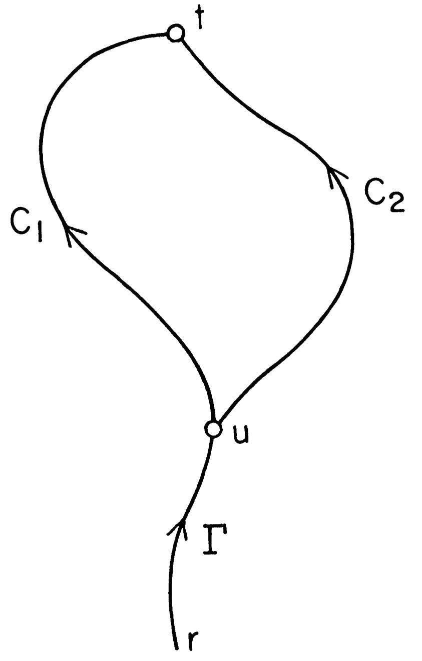


FIG. 1. *An r-lasso.*

If i and j are both basic, then x_i and x_j are strict B-row complements if and only if there is an arc e_k such that e_i and e_j are on the same side of $C(T_B, e_k)$. Letting $e_k = (u, v)$, we have that the set $L = \Gamma(T_B, u) \cup \Gamma(T_B, v) \cup \{e_k\}$ is an r-lasso in $G(A)$, and e_i and e_j are again on the same side of L.

Conversely, suppose that e_i and e_j are on the same side of an r-lasso L in $G(A)$. Then by removing an arc e from the endpoint of L (with $e = e_i$ or e_j if either is an end arc), we obtain an r-tree on the nodes of L. This tree can easily be extended to an r-tree for the entire network for which either $e_j \in C_2(T_B, e_i)$ (if $e = e_i$), $e_i \in C_2(T_B, e_j)$ (if $e = e_j$), or e_i and e_j are on the same side of $C(T_B, e)$ (if $e_i \neq e \neq e_j$). □

Example 4.8. At this point we give an illustration that the various forms of determinacy are not equivalent. Consider the reachability system shown in Fig. 2. Define the vector c by $c_5 = -1$ and $c_j = 1, j \neq 5$. Then $P(c, b)$ satisfies nondegeneracy conditions (N1) and (N2) and has a bounded feasible region. By Theorem 4.7 x_1 and x_5 are $\mathscr{B}_1$-row substitutes, since they occur in only one r-lasso for which they are on opposite sides. However, for the optimal basis $B = \{x_2, x_3, x_4, x_6, x_8\}$—an element of both $\mathscr{B}_1$ and $\mathscr{B}_2$—we have that $e_5 \in C_1(T_B, e_2)$ and $e_1 \in C_2(T_B, e_2)$ so that x_1 and x_5 are strict B-column *complements*. Thus they are not $\mathscr{B}_1$- or $\mathscr{B}_2$-column substitutes. They are, in fact, $\mathscr{B}_1$- and $\mathscr{B}_2$-column complements, and of course $\mathscr{B}_0$-indeterminate. Thus one must be specific as to which of the types of determinacy is meant when referring to variables as substitutes or complements.

Using Theorem 4.7, we can derive a characterization for complete $\mathscr{B}_1$-row determinacy in a reachability model whose feasible region is bounded, or equivalently, where $G(A)$ is acyclic. It is analogous to that found for $\mathscr{B}_0$-determinacy in [P, Thm. 3.7], the seminal result appearing in [D]. Define an *r-rooted Wheatstone bridge on e_i and e_j* to be a subnetwork of the form $\Gamma \cup W$ as shown in Fig. 3, where Γ is a directed path from r to a node w of W which is otherwise disjoint from W. Each path is directed as indicated by the arrows. The nodes t_1 and t_2 may also lie on the path from e_j to v_3, in which case $t_1 = t_2$, and the node w must be on one of the three directed paths from u to e_i, v_2, or t_1/v_3, with the path from w to u directed toward u.

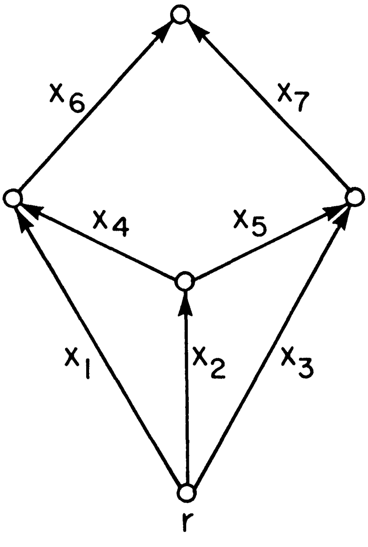


FIG. 2

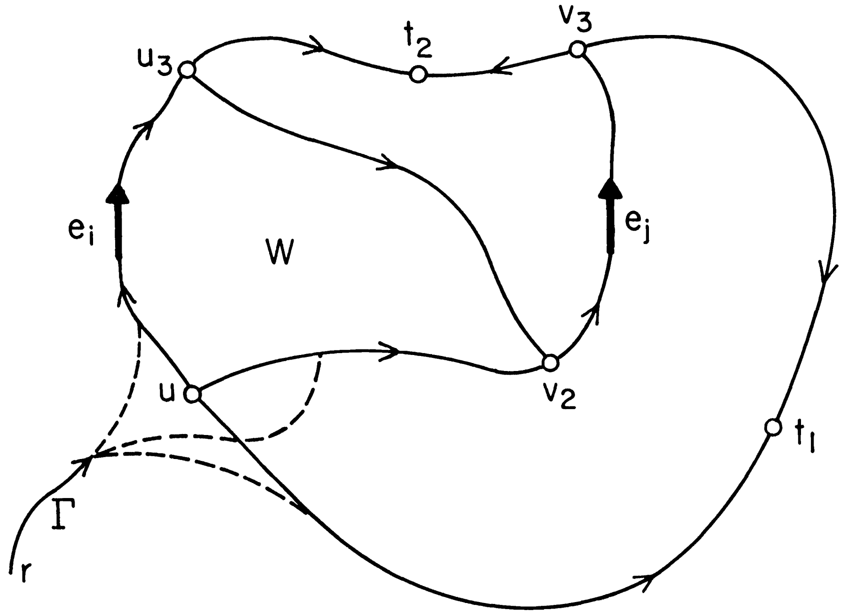


FIG. 3. *An r-rooted Wheatstone bridge.*

THEOREM 4.9. *Let $P(c, b)$ be a reachability model with $G(A)$ acyclic, and x_i and x_j two variables. The x_i and x_j are $\mathscr{B}_1$-row determinate if and only if there is no r-rooted Wheatstone bridge on e_i and e_j contained in $G(A)$.*

Proof. Suppose first that there exists an r-rooted Wheatstone bridge $\Gamma \cup W$ on e_i and e_j. We will take the case when the node w lies on the path from u to e_i, the other cases being similar. Identifying by $[x, y]$ the path in $\Gamma \cup W$ with endpoints x and y, we obtain the two r-rooted lassos

$$L_1 = [r, w] \cup [w, u_3] \cup [u_3, v_2] \cup [v_2, v_3] \cup [v_3, t_1] \cup [w, u] \cup [u, t_1]$$

and

$$L_2 = [r, w] \cup [w, u_3] \cup [u_3, t_2] \cup [w, u] \cup [u, v_2] \cup [v_2, v_3] \cup [v_3, t_2]$$

for which e_i and e_j are on the same side in L_1 and opposite sides in L_2. Therefore by Theorem 4.7 x_i and x_j are indeterminate.

For the converse suppose that x_i and x_j are indeterminate, so that by Theorem 4.7 there exists two r-rooted lassos $L_k = \Gamma_k \cup C_k^1 \cup C_k^2$ with endpoints t_k, $k = 1, 2$ so that e_i and e_j are in the same C_1^l and in opposite C_2^l, $l = 1, 2$. Further choose L_1 and L_2 so that the number of arcs in $L_1 \cup L_2$ is minimized. Relabel e_i, e_j, and the C_k^l so that e_i is in C_2^1, e_j is in C_2^2, and e_i and e_j are in C_1^1, with e_i before e_j. We can consider the L_k to be closed walks, that is, circuits with possibly repeated arcs, and so when we speak of "traversing" L_k, it will be as if it were a circuit. Then L_1 and L_2 can be partitioned into four parts as follows: For L_1, define L_{11} to be that portion of L_1 between r and t_1 which does not contain e_i and e_j, and define L_{12}, L_{13}, and L_{14} to be the portions of L_1 between r and e_i, between e_i and e_j, and between e_j and t_1, respectively. For L_2, define L_{21} and L_{23} to be the portions of L_2 between r and e_i and between e_i and t_2, respectively, and define L_{22} and L_{24} to be the portions of L_2 between r and e_j and between e_j and t_2, respectively. (See Fig. 4.)

CASE 1. Since $G(A)$ is acyclic, L_{21} can intersect L_1 only on $L_{11} \cup L_{12}$. Let u_1 be the farthest node traversing $L_{11} \cup L_{12}$ from e_i at which this occurs, and note that u_1 is on

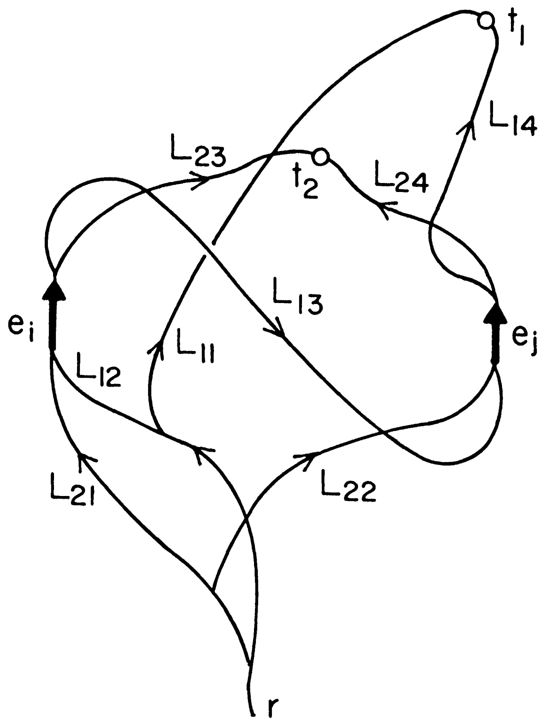


FIG. 4

L_{11} since r is on $L_{11} \cup L_{12}$ and L_{21}. Then that portion of $L_{11} \cup L_{12}$ between e_i and u_1 must be part of L_{21}, since otherwise it could be replaced by the corresponding portion of L_{21} to form lasso L_1' with the same properties of L_1 and whose union with L_2 has fewer arcs, a contradiction. In particular, then, we now have $L_{12} = L_{21}$.

CASE 2. Traverse L_{22} from r to the farthest node u_2 at which L_{22} intersects $L_{11} \cup L_{12}$, and let v_2 be the next node at which L_{22} meets L_1 (u_2 and v_2 must be distinct, since L_{22} begins on $L_{11} \cup L_{12}$ and ends on L_{13}). Suppose that v_2 is on L_{11}. As in Case 1 that portion of L_{11} between u_2 and v_2 can be replaced by the corresponding portion of L_{22}, a contradiction. Thus v_2 must be on L_{13}, since it cannot be on L_{11}, $L_{12} = L_{21}$, or L_{14} (since $G(A)$ is acyclic).

CASE 3. Traverse $L_{23} \cup L_{24}$ from e_i until the farthest node u_3 at which $L_{23} \cup L_{24}$ intersects L_{13}, and let v_3 be the next node at which $L_{23} \cup L_{24}$ again meets L_1 (as in Case 2, u_3 and v_3 must be distinct). Suppose v_3 is on L_{13}. Since $G(A)$ is acyclic, v_3 must also be on L_{23}. Again that portion of L_{13} between u_3 and v_3 can be replaced by the corresponding portion of L_{23}, a contradiction. Thus v_3 must be on $L_{11} \cup L_{14}$ since it cannot be on $L_{12} = L_{21}$.

Cases 1–3 establish

(1) $L_{12} = L_{21}$;

(2) L_{22} coincides with $L_{11} \cup L_{12}$ from r to the point u_2 and its next intersection with L_1 occurs at the point $v_2 \in L_{13}$; and

(3) $L_{23} \cup L_{24}$ coincides with L_{13} from e_i to the point u_3, and its next intersection with L_1 occurs at the point $v_3 \in L_{11} \cup L_{14}$.

(See Fig. 5.) The final parts of L_1 and L_2 to consider are the subpaths P of $L_{11} \cup L_{14}$ and Q of $L_{23} \cup L_{24}$ lying between v_3 and e_j, and the subpaths R of L_{22} and S of L_{13} lying between v_2 and e_j. From the statements above it follows that P and S intersect L_2 only

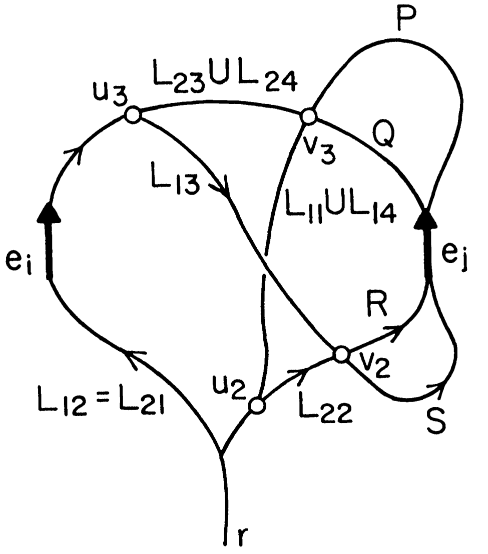


FIG. 5

on Q and R, and Q and R intersect L_1 only on P and S. We claim that $P = Q$ and $R = S$, since if v_3 is on L_{24} then P and S can be replaced by Q and R, if v_3 is on L_{23} then Q and R can be replaced by P and S, in each case forming lassos L'_1 and L'_2 having smaller union than L_1 and L_2, a contradiction.

The situation is now as shown in Fig. 6, with either $t_1 = t_2$ on $\Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ or t_1 on Γ_1, and t_2 on Γ_2, and Γ_2 ending on $\Gamma_4 \cup \Gamma_5 \cup \Gamma_6$. This forms an r-rooted Wheatstone bridge as given in Fig. 3, with $u = u_1$ and $w = u_2$ on the path from u to v_2 if Γ_2 ends at Γ_4, $u = u_2$ and $w = u_1$ on the path from u to e_i if Γ_2 ends at Γ_5, and $u = u_2$ and $w = u_1$ on the path from u to v_3 if Γ_2 ends at Γ_6. The theorem follows. □

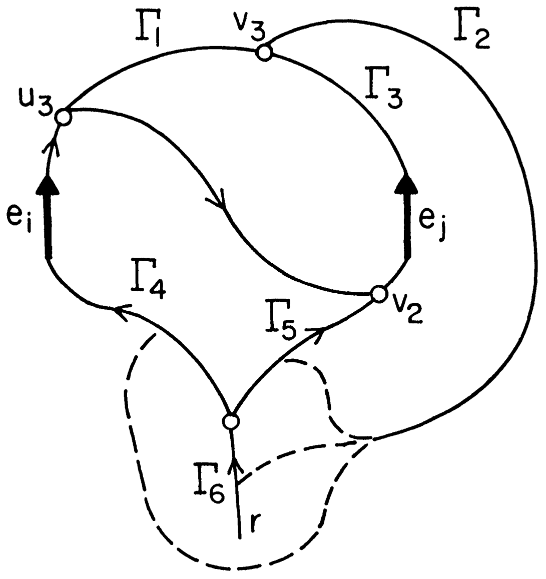


FIG. 6

The restriction that $G(A)$ be acyclic is essential to Theorem 4.9. There exists pathological nonacyclic networks for which two variables are not determinate, but for which no easily characterizable Wheatstone-type subgraph seems to exist. The problem of characterizing determinate pairs of variables in a general reachability model is therefore still an open problem.

The final result concerns determinacy in a special class of reachability models. A transshipment matrix A is called a *transportation matrix* if the rows of A can be partitioned into subsets U and V so that

(1) if a column of A has a single nonzero entry, then this entry occurs in a row of U if it is $+1$, and in a row of V if it is -1;

(2) if a column of A has two nonzero entries, then the -1 entry occurs in a row of U and the $+1$ entry occurs in a row of V;

(3) for each row i of A there is column of A with a $+1$ entry in row i.

Assumption (3) insures that $G(A)$ is r-connected. Associated with an extended transportation matrix is a bipartite network $G'(A)$, consisting of $G(A)$ with node r and its adjacent arcs removed, that all arcs are members of $U \times V$. The variables whose columns contain only $+1$ are called *supply* variables, those whose columns contain only -1 are called *demand* variables, and those whose columns contain both a $+1$ and a -1 are called *transportation* variables. Supply variables will be denoted by $s(u)$, demand variables by $d(v)$, and transportation variables by $t(u, v)$, where u, v, or (u, v) is the node or arc associated with that variable. Since multiple copies of a variable are clearly substitutes with each other and identical in their relationship to the other variables, we assume that there is only one copy of each variable. Further, if there is any node having exactly one adjacent arc of $G(A)$, then the associated variable will be independent of all other variables and will have no effect on determinacy in the system. We therefore assume all such variables (and associated nodes) are removed.

Determinacy has been characterized for transportation models over the basis system $\mathscr{B}_0$ in Theorem 3.9 of [P]. It turns out that for the system $\mathscr{B}_1$ there is also an easy characterization of determinacy.

THEOREM 4.10. *Let $P(c, b)$ be a reachability model associated with a transportation matrix. Then every pair of variables is $\mathscr{B}_1$-row determinate. In particular, in terms of the associated bipartite network $G'(A)$ we have*:

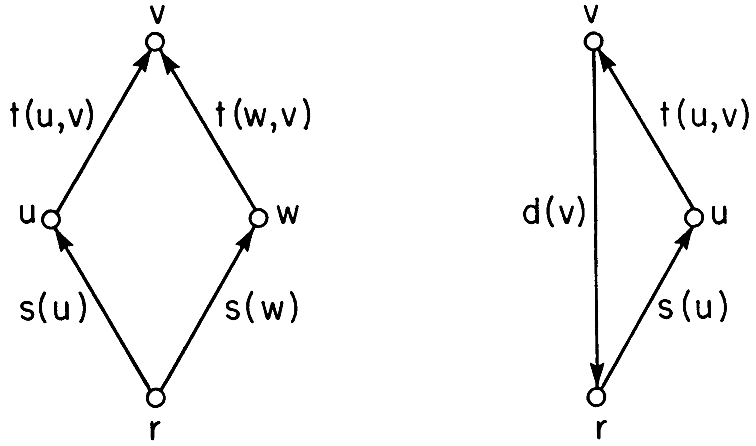


FIG. 7. *r-lassos in the transportation model.*

(i) *variables $s(u)$ and $s(w)$ are strict substitutes if u and w are adjacent to the same node, and independent otherwise*;

(ii) *variables $d(u)$ and $d(z)$ are independent*;

(iii) *variables $s(u)$ and $d(v)$ are strict complements if (u, v) is an arc, and independent otherwise*;

(iv) *variables $s(u)$ and $t(w, v)$ are strict complements if $u = w$, strict substitutes if $u \neq w$* and *(u, v) is an arc, and independent otherwise*;

(v) *variables $d(v)$ and $t(w, z)$ are strict complements if $v = z$, and independent otherwise*;

(vi) *variables $t(u, v)$ and $t(w, z)$ are strict substitutes if $v = z$, and independent otherwise.*

Proof. Using Theorem 4.7, we look at all r-lassos in $G(A)$, and note that there are only two classes of r-lassos, which are shown in Fig. 7. By checking the position of supply, demand, and transportation variables on these r-lassos, and noting that there are no nodes of V which have only one adjacent arc, it follows that the only cases when two variables are strict B-row substitutes or strict B-complements are those given by the theorem.

Acknowledgments. The author would like to acknowledge helpful discussions with A. S. Kydes and H. V. Tauchen, and comments by the referees of earlier versions of the manuscript.

REFERENCES

[D] R. J. Duffin, *Topology of series-parallel networks*, J. Math. Anal. Appl., 10 (1965), pp. 303–318.

[GP] D. Gale and T. Politof, *Substitutes and complements in network flow problems*, Disc. Appl. Math., 3 (1981), pp. 175–186.

[Go] T. Gorman, *More scope for qualitative economics*, Rev. Econom. Stud., 31 (1964), pp. 65–68.

[GV] F. Granot and A. F. Veinott, *Substitutes, complements and ripples in network flows*, Math. Oper. Res., 10 (1985), pp. 471–497.

[Gr] H. J. Greenberg, *Measuring complementarity and qualitative determinacy in matricial forms*, in Computer-Assisted Analysis and Model Simplification, Associated Press, New York, 1981, pp. 497–522.

[GLM] H. J. Greenberg, J. R. Lundgren and J. S. Maybee, *Rectangular matrices and signed graphs*, this Journal, 4 (1983), pp. 50–61.

[GM] H. J. Greenberg and J. S. Maybee, *Computer-Assisted Analysis and Model Simplification*, Associated Press, New York, 1981.

[H] J. R. Hicks, *Value and Capital*, 2nd Edition, Oxford University Press, Clarendon, Oxford, 1939.

[HA] J. R. Hicks and R. G. D. Allen, *A reconsideration of the theory of value*, Parts I and II, Economica, 1 (1934), pp. 52–76, 196–219.

[J] C. R. Johnson, *Sign patterns of inverse nonnegative matrices*, Linear Algebra Appl., 55 (1983), pp. 69–80.

[JLR] C. R. Johnson, F. T. Leighton and H. A. Robinson, *Sign patterns of inverse-positive matrices*, Linear Algebra Appl., 24 (1979), pp. 75–83.

[KL] V. Klee and R. Ladner, *Qualitative matrices: strong sign-solvability and weak satisfiability*, in Computer-Assisted Analysis and Model Simplification, Associated Press, New York, 1981, pp. 293–320.

[KLM] V. Klee, R. Ladner and R. Manber, *Sign solvability Revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.

[KWW] G. J. Koehler, A. B. Whinston and G. P. Wright, *Optimization on Leontief Substitution Systems*, North-Holland, Amsterdam, 1975.

[La] G. M. Lady, *The structure of qualitative relationships*, Econometrica, 51 (1983), pp. 197–218.

[L1] K. Lancaster, *The scope of qualitative economics*, Rev. Econom. Stud., 29 (1962), pp. 99–132.

[L2] ———, *The solution of qualitative comparative static problems*, Quart. J. Econom., 80 (1966), pp. 278–295.

[L3] K. LANCASTER, *The theory of qualitative linear systems*, Econometrica, 33 (1965), pp. 395–408.

[M] R. MANBER, *Graph theoretical approach to qualitative solvability of linear systems*, Linear Algebra Appl., 48 (1982), pp. 457–470.

[MQ] J. MAYBEE AND J. QUIRK, *Qualitative problems in matrix theory*, SIAM Rev., 11 (1969), pp. 30–51.

[Mu] K. MURTY, *Linear Programming*, John Wiley, New York, 1983.

[My] J. MAYBEE, *Sign solvability*, in Computer-Assisted Analysis and Model Simplification, Associated Press, New York, 1981, pp. 201–258.

[P] J. S. PROVAN, *Determinacy in linear systems and networks*, this Journal, 4 (1983), pp. 262–278.

[PK] J. S . PROVAN AND A. S. KYDES, *Correlation and determinacy in network models*, BNL Rep. 51243, Brookhaven National Laboratory, Upton, NY, 1980.

[R] G. RITSCHARD, *Computable qualitative comparative static techniques*, Econometrica, 51 (1983), pp. 1145–1168.

[S1] P. A. SAMUELSON, *Foundations of Economic Analysis*, Anthenum, New York, 1971 (originally published in 1947 by Harvard University Press, Cambridge, MA).

[S2] ———, *Complementarity: An essay on the fortieth anniversary of the Hicks–Allen revolution in demand theory*, J. Econom. Lit., (1974), pp. 1255–1289.

[Sh1] L. S. SHAPLEY, *On network flow functions*, Naval Res. Logist. Quart., 8 (1961), pp. 151–158.

[Sh2] ———, *Complements and substitutes in the assignment problem*, Naval Res. Logist. Quart., 9 (1962), pp. 45–48.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

007

ON THE COVERING RADIUS PROBLEM FOR CODES I. BOUNDS ON NORMALIZED COVERING RADIUS*

KAREN E. KILBY† AND N. J. A. SLOANE‡

Abstract. In this two-part paper we introduce the notion of a stable code and give a new upper bound on the normalized covering radius of a code. The main results are that, for fixed k and large n, the minimal covering radius $t[n, k]$ is realized by a normal code in which all but one of the columns have multiplicity 1; hence $t[n + 2, k] = t[n, k] + 1$ for sufficiently large n. We also show that codes with $n \leqq 14$, $k \leqq 5$ or $d_{\min} \leqq 5$ are normal, and we determine the covering radius of all proper codes of dimension $k \leqq 5$. Examples of abnormal nonlinear codes are given. In Part I we investigate the general theory of normalized covering radius, while in Part II [this Journal, 8 (1987), pp. 619-627] we study codes of dimension $k \leqq 5$, and normal and abnormal codes.

Key words. binary codes, covering radius

AMS(MOS) subject classifications. 05B, 94B

1. Introduction. Let C be an $[n, k]$ binary linear code. The covering radius R (also denoted by $CR(C)$) is given by

$$R = \max_{x \in \mathbf{F}^n} \min_{c \in C} d(x, c),$$

where $\mathbf{F} = \{0, 1\}$ and $d(\ ,\ )$ is Hamming distance. Let $t[n, k]$ denote the smallest R for any $[n, k]$ code. Two central problems in this subject are to determine $t[n, k]$ and to construct codes with $R = t[n, k]$ (see [1]-[3], [9], [10] for further background information).

Before describing the new results, we define the normalized covering radius, which as we shall see is easier to work with than the covering radius itself. Let C have generator matrix G. In general, G may contain repeated columns. *We assume throughout, however, that no column of G is zero.* Let a be the number of distinct columns occurring in G, and let $m_1, \cdots, m_a$ be their multiplicities, with $m_1 + \cdots + m_a = n$. Then

$$R \geqq \sum_{i=1}^{a} \left[\frac{m_i}{2}\right],$$

and, following [10], we define the *normalized covering radius* ρ of C to be

$$\rho = R - \sum_{i=1}^{a} \left[\frac{m_i}{2}\right], \tag{1}$$

a nonnegative integer. Then

$$R = \sum_{i=1}^{a} \left[\frac{m_i}{2}\right] + \rho = \frac{n}{2} - \frac{\text{no. of odd } m_i}{2} + \rho. \tag{2}$$

Summary of results. A *stable code* (§ 3) has the property that ρ does not increase when any number of pairs of identical columns of any length are adjoined to it. Many small codes are stable (§ 6 of Part II), so this often provides a quick method for determining the covering radius. The *contracted* code $\tilde{C}$ (§ 3) is spanned by the rows of the matrix formed by taking one copy of each column of G that has odd multiplicity, where G is

* Received by the editors February 6, 1986; accepted for publication (in revised form) March 24, 1987.

† Morse College, Yale University, New Haven, Connecticut 06520. The work of this author was supported in part by Bell Laboratories.

‡ Mathematical Sciences Research Center, AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

a generator matrix for C. If $\tilde{C}$ is stable, $\rho(C) = \rho(\tilde{C})$. As an illustration, consider the code C with generator matrix

$$\begin{bmatrix} 111 & 000 & 00 & 00 & 11 \\ 000 & 111 & 00 & 00 & 11 \\ 000 & 000 & 11 & 00 & 11 \\ 000 & 000 & 00 & 11 & 11 \end{bmatrix}$$

(with multiplicities 3, 3, 2, 2, 2), encountered in the proof of Theorem 27 of [2]. The contracted code $\tilde{C}$ has generator matrix $\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$, and is stable with $\rho = 0$, so C has covering radius $0 + \sum [m_i/2] = 5$. Stable codes are normal (see Theorem 4).

Theorems 6 and 7 give improved upper bounds on ρ. Section 5 considers how ρ increases when the multiplicities of the columns are increased, subject to the constraint that the *parities* of the multiplicities are unchanged. More precisely, fix an $[n_B, k]$ *projective* code B (i.e., one with distinct columns), and consider all $[n, k]$ codes C with $\tilde{C} = B$. For sufficiently large n, $\rho_\infty(B) = \max_C \rho(C)$ and $\rho_*(B) = \min_C \rho(C)$ are independent of n. Theorem 8 investigates how rapidly $\rho_\infty(B)$ can be reached. Theorem 9 shows that $\rho_*(B)$ can be realized by a normal code having a very special structure, in which all columns have multiplicity 1 except for one column that has large multiplicity. Furthermore a normal code C has $\rho_*(C) = \rho(C)$ (see Theorem 11).

For fixed k and large n, the minimal covering radius of any $[n, k]$ code is given by

$$t[n,k] = \frac{n}{2} + \min_B \left\{ \rho_*(B) - \frac{n_B}{2} \right\},$$

where B ranges over the projective codes of dimension k or $k - 1$ (equation (52)). It follows (see Theorem 12) that, for large n, $t[n, k]$ can be attained by a normal code having the above-mentioned special structure. This establishes Conjectures A and D of [2] for sufficiently large n. A heuristic justification for the special structure of these codes is given at the end of § 6 of Part II.

Codes of dimension $k \leqq 4$ were studied in [10]. We have now determined the covering radius of every projective code of dimension 5. If C is any $[n, 5]$ code, then

$$CR(\tilde{C}) \leqq \rho(C) \leqq CR(\tilde{C}) + 1$$

(see Theorem 13 of Part II). This implies that all codes of dimension $\leqq 5$ are normal.

Table 2(a) of Part II gives upper and lower bounds on ρ, and enables one to write down the covering radius of any code C for which the contracted code $\tilde{C}$ has dimension $\leqq 5$, with an error of at most 1, when only the length and dimension of $\tilde{C}$ are known. For example, suppose C is a [3000, 12] code for which $\tilde{C}$ is a [20, 5] code. From Table 2(a) of Part II we see that $7 \leqq \rho(C) \leqq 9$, or in other words (using (2))

$$CR(C) = \frac{3000}{2} - \frac{20}{2} + 8 + \theta = 1498 + \theta,$$

where $\theta = -1, 0$ or 1.

Section 6 presents a summary of the projective codes of dimension 5 and length $5 \leqq n \leqq 31$, and gives one or two examples of the best covering codes of each length (see especially Fig. 3 of Part II). This list of codes should be useful, since the investigation of the subject has been hampered by a shortage of examples of good covering codes. The precise determination of ρ for some of these codes requires a separate analysis, as illustrated in Theorem 10, and we have only carried this out in certain cases. At the end of § 6 of Part II we construct an infinite family of (normal) codes with unacceptable coordinates.

Sections 7 and 8 of Part II show that if $d_{\min} \leqq 5$ or if $\tilde{C}$ has $d_{\min} \leqq 2$ then C is normal. Theorem 18 of Part II summarizes the known conditions on the parameters of a code that imply normality. Finally, § 9 of Part II gives Peter Frankl's construction of an abnormal nonlinear code.

Definitions and notation. The minimal distance of a code is denoted by $d_{\min}$, and the order of its automorphism group by g. We use $\mathbf{F}$ to denote both the Galois field $GF(2)$ and the code $\{0, 1\}$. The empty code of length zero will be denoted by 0, and E_n denotes the $[n, n-1]R = 1$ *even weight* code. The $[n, 1]R = [\frac{1}{2}n]$ *repetition* code T_n contains 0^n and 1^n. The $[n = 2^k - 1, k]R = 2^{k-1}$ *simplex* code S_k is defined by a generator matrix in which the columns comprise all distinct nonzero binary k-tuples (see (14) and Fig. 3 of Part II). In particular, $S_0 = 0$, $S_1 = \mathbf{F}$, $S_2 = E_3$. The $2^k - 1$ columns of a generator matrix for S_k, for $k \geqq 3$, may be regarded as representing the points of a projective geometry $PG(2, k-1)$ of dimension $k-1$ over $\mathbf{F}$. In such a geometry every line contains exactly three points; three points are collinear if and only if the corresponding vectors sum to zero. We shall occasionally use this geometrical language even when k is less than 3.

Normal codes. Let C be a linear or nonlinear code of length n and covering radius R. For $i = 1, \cdots, n$ and $a = 0, 1$ let $C_a^{(i)}$ denote the subset of codewords $(c_1, \cdots, c_n)$ of C with $c_i = a$, and for an arbitrary $x \in \mathbf{F}^n$ let

$$f_a^{(i)}(x) = d(x, C_a^{(i)}),$$

if $C_a^{(i)}$ is nonempty, and let $f_a^{(i)}(x) = n$ otherwise. Then

$$N^{(i)} = \max_x \{f_0^{(i)}(x) + f_1^{(i)}(x)\}$$

is called the *norm of C with respect to the ith coordinate.* If

$$N^{(i)} \leqq N \tag{3}$$

for at least one coordinate i, we say that C has *norm*[1] N, and coordinates i for which (3) holds are called *acceptable*, the other coordinates being *unacceptable.* Finally, C is *normal* if it has norm N satisfying

$$N \leqq 2R + 1, \tag{4}$$

and is otherwise *abnormal.* It follows from the definition that if C has norm N, it also has norm $N+1, N+2, \cdots$. We take N as small as possible. For any code,

$$2R \leqq N. \tag{5}$$

Many other properties of the norm will be found in [3].

2. Normalized covering radius ρ. Let C be an $[n, k]R$ code (assumed throughout to have no coordinate position that is identically zero). Then [1, Thm. 6]

$$R \leqq \left[\frac{n}{2}\right]. \tag{6}$$

The normalized covering radius $\rho(C)$, defined in (1), satisfies (see [10]) $\rho(C) \geqq 0$; if all m_i are even, $\rho(C) = 0$; if all m_i are 1, $\rho(C) = R$. (Note that $\rho(C)$ does not depend on the

[1] The meanings of three terms in this paper differ from those used in certain earlier works, as will be indicated in footnotes. The new definitions seem preferable. This definition of norm is the one used in [2], [10] and the manuscript of [3]. The definition given in the published version of [3] turns out to be less satisfactory.

choice of a generator matrix.) By reordering the coordinates (if necessary) we may assume that the first m_1 columns of C are identical, then the next m_2 columns, and so on. We partition the codewords $c \in C$ as

$$(7) \qquad c=(c^{(1)}, c^{(2)}, \cdots, c^{(a)}), \qquad c^{(i)}=(c_i, c_i, \cdots, c_i)$$

where length $(c^{(i)}) = m_i$. Correspondingly, we partition an arbitrary vector $x \in \mathbf{F}_2^n$ as

$$(8) \qquad x=(x^{(1)}, \cdots, x^{(a)}),$$

where length $(x^{(i)}) = m_i$. The *height of* $x^{(i)}$ is defined to be

$$(9) \qquad h_i = ht(x^{(i)}) = wt(x^{(i)}) - \left[\frac{m_i}{2}\right],$$

and the *height of* x is

$$(10) \qquad ht(x) = \sum_{i=1}^{a} h_i.$$

Then we have ([10, (15)])

$$(11) \qquad \rho(C) = \max_x \min_{c \in C} ht(x+c).$$

A vector x such that $\min_{c \in C} ht(x + c) = \rho(C)$, or equivalently $d(x, C) = R$, is called a *deep hole* in C. It is shown in Theorem 1 of [10] that, using (11), we can express the problem of finding $\rho(C)$ as an integer programming problem.

The following result will turn out to be very useful.

THEOREM 1. *Suppose a code C is the row space of a matrix of the form* $\left[\begin{smallmatrix} G_1 & G_3 \\ 0 & G_2 \end{smallmatrix}\right]$, *where G_1 and G_2 have no columns of* 0*'s, and let codes A and B be the row spaces of G_1 and G_2, respectively.* (a) *If $G_3 = 0$, so that C is a direct sum $C = A \oplus B$, then*

$$(12) \qquad \rho(C) = \rho(A) + \rho(B),$$

and if either A or B is normal so is C. (b) *If all columns in* $\left[\begin{smallmatrix} G_3 \\ G_2 \end{smallmatrix}\right]$ *occur an even number of times, then $\rho(C) = \rho(A)$, and if A is normal so is C.*

Proof. (a) Clearly

$$(13) \qquad CR(C) \leqq CR(A) + CR(B)$$

(cf. [7]), with equality if $G_3 = 0$, which implies (12). Suppose $G_3 = 0$, A is normal and coordinate r is acceptable. Let A, B, C have lengths n_A, n_B, n_C, respectively. For an arbitrary $x = |y|z| \in \mathbf{F}^{n_C}$, $y \in \mathbf{F}^{n_A}$, $z \in \mathbf{F}^{n_B}$ we have, in the notation of § 1,

$$d(x, C_a^{(r)}) = d(y, A_a^{(r)}) + d(z, B), \qquad a = 0, 1,$$

$$\begin{aligned} \text{Norm }(C) &= \text{Norm }(A) + 2CR(B) \\ &\leqq 2CR(A) + 1 + 2CR(B) \\ &= 2CR(C) + 1, \end{aligned}$$

and C is normal. (b) Now $CR(C) = CR(A) + \frac{1}{2}n_B$, so $\rho(C) = \rho(A)$. If A is normal, then $d(x, C_a^{(r)}) \leqq d(y, A_a^{(r)}) + \frac{1}{2}n_B$, and again Norm $(C) \leqq 2CR(C) + 1$.

3. The effect on ρ of varying the multiplicities; stable codes. We now investigate how $\rho(C)$ changes as the multiplicities m_i of the columns of C vary. Any code C of dimension $\leqq k$ can be obtained by assigning suitable multiplicities to the columns of the simplex code S_k. Let us arrange the columns of S_k in some fixed order, for example, the binary order illustrated in (14) for S_3 (see also Fig. 3 of Part II):

$$S_3 : \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{14}$$

We choose a generator matrix G for C, and let $m_i \geqq 0$ be the number of times the ith column of S_k appears in G, for $i = 1, \cdots, 2^k - 1$. Then we let

$$\rho^{(k)}(m_1, m_2, \cdots, m_{2^k-1})$$

equal the normalized covering radius $\rho(C)$. Choosing a different generator matrix for C permutes the m_i's but does not change the value of $\rho^{(k)}(m_1, \cdots, m_{2^k-1})$. (However most permutations of the arguments do change the value of $\rho^{(k)}(m_1, \cdots, m_{2^k-1})$.)

It turns out that the function $\rho^{(k)}$ is best studied by allowing the m_i to vary while fixing their *parity*, or in other words by investigating how $\rho^{(k)}$ changes when pairs of identical columns are added to or deleted from the generator matrix G. It is an elementary fact that when two identical columns are adjoined to G (columns which may or may not already be present in G), the covering radius of C increases by either 1 or 2, so the normalized covering radius is either unchanged or increases by 1. This establishes the *monotonicity property* [10, Thm. 2]: if $m_i \leqq m'_i$ and $m_i \equiv m'_i$ (mod 2) for all i, then

$$\rho^{(k)}(m_1, \cdots, m_{2^k-1}) \leqq \rho^{(k)}(m'_1, \cdots, m'_{2^k-1}). \tag{15}$$

The earliest codes to be considered are therefore the projective codes. Given an arbitrary code C, with parameters $[n, k]R$ and generator matrix G, the corresponding *contracted code*[2] $\tilde{C}$ is the projective code which is the row space of the matrix formed by taking one copy of each column of G that has odd multiplicity. $\tilde{C}$ is independent of the choice of G. (If all m_i are even we set $\tilde{C}$ equal to the empty code 0.) We denote the parameters of $\tilde{C}$ by $[\tilde{n}, \tilde{k}]\tilde{R}$, so $\tilde{n} \leqq n$, $\tilde{k} \leqq k$, $\tilde{R} \leqq R$ and $\rho(\tilde{C}) \leqq \rho(C)$.

We say that two codes C, D are *congruent* (written $C \equiv D$) if $\tilde{C} = \tilde{D}$. If the multiplicities of the columns in C do not exceed the multiplicities of the same columns in D (but with no constraints on their parities), we write $C \leqq D$. For example, the contracted code $\tilde{C}$ satisfies

$$\tilde{C} \equiv C \quad \text{and} \quad \tilde{C} \leqq C. \tag{16}$$

The monotonicity property (15) states that

$$C \equiv D, \quad C \leqq D \quad \Rightarrow \quad \rho(C) \leqq \rho(D). \tag{17}$$

We shall need the following corollary to Theorem 19 of [3].

THEOREM 2. *Let C be normal and suppose the rth coordinate is acceptable. Let D be formed by adjoining $2m$ copies of the rth coordinate to C. Then $CR(D) = CR(C) + m$,* Norm $(D) =$ Norm $(C) + 2m$, *D is normal, and any copy of the rth coordinate is an acceptable coordinate for D.*

Proof. D is an amalgamated direct sum [3] of C with the repetition code T_{2m+1}. By the remarks made earlier in this section, $CR(D) \geqq CR(C) + m$. From Theorem 19(ii) of [3], $CR(D) = CR(C) + m$. Then D is normal by Theorem 19(iii) of [3].

The determination of the covering radius of codes of low dimension is greatly facilitated by the observation that for many of these codes ρ does not increase when pairs of identical columns are adjoined to the generator matrix. We call C *stable* if it has this

[2] A different definition of contracted code was used in [2].

property, or more precisely if

$$C \equiv D, \quad C \leqq D \quad \Rightarrow \quad \rho(C) = \rho(D). \tag{18}$$

We shall see, for example, that all codes of dimension $k \leqq 3$ are stable.

In view of Theorem 1(b), adjoining pairs of identical columns from outside the subspace of $PG(2, k-1)$ spanned by the columns of C has no effect on ρ, so such columns can be ignored when investigating the stability of C.

THEOREM 3. *Let an* $[n, k]$ *code* C *be the row space of a matrix* G, *and let* Π *be the subspace of* $PG(2, k-1)$ *spanned by the columns of* G. *Then* C *is stable if and only if* $\rho(C)$ *does not increase when any number of pairs of identical columns representing points in* Π *are adjoined to* G.

Examples. If all the multiplicities m_i are even, then $\rho = 0$. But $\tilde{C}$ is the empty code 0, with $\rho = 0$. We deduce that 0 is stable. (This can also be deduced directly from the theorem.)

If C has dimension 1 then (since no coordinate may be identically zero) $C = T_n$, $R = [n/2]$, $\rho = 0$. But $\tilde{T}_n = 0$ if n is even, $\tilde{T}_n = \mathbf{F}$ if n is odd, both having $\rho = 0$. We deduce that $\mathbf{F}$ is stable. In fact all the codes $\mathbf{F}^n$ and E_n are stable (see the examples preceding Theorem 6).

Any code C of dimension 2 has a generator matrix containing (say) a columns $\binom{0}{1}$, b columns $\binom{1}{0}$ and c columns $\binom{1}{1}$. As stated on page 388 of [3], the covering radius of C is given by

$$\left[\frac{a}{2}\right] + \left[\frac{b}{2}\right] + \left[\frac{c}{2}\right] + 1 \quad \text{if } a, b, c \text{ are odd,} \tag{19}$$

$$\left[\frac{a}{2}\right] + \left[\frac{b}{2}\right] + \left[\frac{c}{2}\right] \quad \text{otherwise.} \tag{20}$$

This now has a very short proof. We calculate $\tilde{C}$, which is

$$\begin{aligned} &0, \text{ with } \rho = 0, && \text{if } a, b, c \text{ are even,} \\ &\mathbf{F}, \text{ with } \rho = 0, && \text{if one of } a, b, c \text{ is odd,} \\ &\mathbf{F}^2, \text{ with } \rho = 0, && \text{if two of } a, b, c \text{ are odd,} \\ &E_3, \text{ with } \rho = 1, && \text{if } a, b, c \text{ are odd.} \end{aligned}$$

All four codes are stable, and (19), (20) follow immediately.

THEOREM 4. *A stable code is normal, and all coordinates are acceptable.*

LEMMA 5. *Let* C *be any* $[n, k]$ *code such that for some* i $(1 \leqq i \leqq n)$, *and all* $l = 0, 1, 2, \cdots$, *adjoining* $2l$ *copies of column* i *to* C *does not increase the normalized covering radius* ρ. *Then* C *is normal and coordinate* i *is acceptable.*

Proof. Suppose coordinate i is unacceptable. Therefore there is a vector x such that $f_0^{(i)}(x) + f_1^{(i)}(x) \geqq 2R + 2$. Without loss of generality $f_0^{(i)}(x) \leqq f_1^{(i)}(x)$, say $f_0^{(i)}(x) = R - \theta$, $f_1^{(i)}(x) \geqq R + \theta + 2$, where $0 \leqq \theta \leqq R$. We construct D by adjoining $2R + 2$ copies of column i to C. Then $\rho(D) = \rho(C)$, so

$$CR(D) = CR(C) + R + 1 = 2R + 1. \tag{21}$$

Let $x^* = |x|u|$, where u is a vector of length $2R + 2$ and weight $w = R + \theta + 2$. Then for D,

$$f_0^{(i)}(x^*) = f_0^{(i)}(x) + w = 2R + 2,$$

$$f_1^{(i)}(x^*) = f_1^{(i)}(x) + 2R + 2 - w \geqq 2R + 2,$$

which contradicts (21).

Theorem 4 now follows because a stable code satisfies the hypothesis of the lemma for every i.

4. Upper bounds on ρ. Suppose C is an $[n, k]R$ code, and the contracted code $\tilde{C}$ is an $[\tilde{n}, \tilde{k}]\tilde{R}$ code. It follows immediately from (1), (6) that

$$\rho(C) \leqq \left[\frac{\tilde{n}}{2}\right]. \tag{22}$$

A stronger upper bound was given in Theorem 8 of [10]. The main goal of the present section is to prove Theorem 7, which strengthens both of these results. We first quote (from Corollary 6 of [7]) the result that the *maximal* covering radius of any $[n, k]$ code is given by

$$T[n,k] = \begin{cases} \left[\dfrac{n}{2}\right] & \text{for} \quad 1 \leqq k \leqq \left[\dfrac{n}{2}\right], \qquad (23) \\[2ex] n-k & \text{for} \quad \left\lceil\dfrac{n}{2}\right\rceil \leqq k \leqq n \qquad (24) \end{cases}$$

where $\lceil x \rceil$ denotes the smallest integer $\geqq x$.

The method of *pivoting*, introduced in § VII of [10], is a useful technique for getting upper bounds on $\rho(C)$ (which are often tight), and leads to Theorem 6. Consider C to be formed from the simplex code S_k with appropriate multiplicities m_i, with length $n = \sum m_i$. We partition vectors of $\mathbf{F}^n$ into blocks as in (7), (8). We choose a coordinate $Q(1 \leqq Q \leqq 2^k - 1)$, called the *pivot*, such that $m_Q \neq 0$. For an arbitrary vector x we first make $ht(x^{(Q)}) \leqq 0$ by (if necessary) adding a codeword $c = (c^{(1)}, \cdots, c^{(2^k-1)}) \in C$ for which $c^{(Q)} \neq 0$.

Let $C_a^{[Q]}$ denote the set of all codewords of C for which $c^{(Q)} = a$, with the Qth block of coordinates deleted, for $a = 0, 1$. $C_0^{[Q]}$ is a code of length $n - m_Q$ and dimension $k - 1$. $C_1^{[Q]}$ is a translate of $C_0^{[Q]}$ and has the same covering radius. In particular, $C_0^{[Q]}$ is obtained by assigning multiplicities m'_P (say) to S_{k-1}. The m'_P are related to the original multiplicities m_P as follows. The $m_P(1 \leqq P \leqq 2^k - 1)$ are nonnegative integers assigned to the points $P \in PG(2, k-1)$. When we form the subcode $C_0^{[Q]}$, the m_P are combined in pairs to give the new multiplicities m'_P. The multiplicities m_R and m_S are combined if and only if QRS is a line in $PG(2, k-1)$. Thus

$$m'_P = m_R + m_S \quad \text{for } QRS \text{ a line in } PG(2, k-1). \tag{25}$$

In particular, the number of distinct columns in $C_0^{[Q]}$ with odd multiplicity, ν say, is equal to the number of lines QTU for which one of m_T and m_U is odd and the other even.

We return to the problem of reducing the distance from x to C. By adding a suitable codeword of $C_0^{[Q]}$ we can make $ht(x) \leqq \rho^{(k-1)}(m'_1, \cdots, m'_{2^{k-1}-1})$. This leads to the *pivoting bound* [10, Thm. 7]: if $m_Q \neq 0$,

$$\rho(C) \leqq \eta + \rho(C_0^{[Q]}), \tag{26}$$

i.e.,

$$\rho^{(k)}(m_1, \cdots, m_{2^k-1}) \leqq \eta + \rho^{(k-1)}(m'_1, \cdots, m'_{2^{k-1}-1}), \tag{27}$$

where η is the number of lines QRS for which m_R and m_S are odd, and the m'_P are given by (25). In the other direction, we have (from (17))

$$\rho(\tilde{C}) \leqq \rho(C). \tag{28}$$

It is often convenient to be able to refer directly to the code obtained by contracting $C_0^{[Q]}$, which we shall denote by $\hat{C}$. In particular, if $\hat{C}$ is stable, we have

$$\rho(C) \leqq \eta + \rho(\hat{C}). \tag{29}$$

Remarks. (1) Different choices for the pivot Q may give different bounds, so we can replace the right side of (27) by

$$\min_Q \{\eta + \rho^{(k-1)}(m'_1, \cdots, m'_{2^{k-1}-1})\}. \tag{30}$$

It appears best to choose Q so that m_Q is odd. Even so, (30) may not be tight: there may be no Q for which equality holds in (27). If there is such a Q we call C and $\tilde{C}$ *tame*, otherwise *wild*. (2) It was shown in [10] (see especially Theorem 9) that if dim C = dim $\tilde{C}$ and the bounds (26) and (28) differ by at most 1, i.e., if, for some choice of the pivot Q, either $\rho(\tilde{C}) = \eta + \rho(C_0^{[Q]})$, or $\rho(\tilde{C}) = \eta + \rho(C_0^{[Q]}) - 1$, then C is normal. (It is easy to show in fact that the hypothesis dim C = dim $\tilde{C}$ is unnecessary.)

Examples. Suppose C is an $[n, k]$ code such that $\tilde{C} = \mathbf{F}^k$. We show by induction on k that $\rho(C) = 0$. Without loss of generality C has a generator matrix $[IG_1]$ where I is an identity matrix and all the columns of G_1 have even multiplicity. By pivoting on the first coordinate, we obtain $\rho(C) = 0$. Therefore $\mathbf{F}^k$ is stable for all k, and $\rho(F^k) = 0$. A similar argument shows that E_k is stable and $\rho(E_2) = 0$, $\rho(E_k) = 1$ for $k \geqq 3$.

THEOREM 6. *Let C be an $[n, k]$ code and let the contracted code $\tilde{C}$ be an $[\tilde{n}, \tilde{k}]$ code. Then*

$$\rho(C) \leqq T[\tilde{n}, \tilde{k}]. \tag{31}$$

Proof. The proof is by induction on k, the result being immediate for $k = 0$ and 1. Let C have multiplicities m_P, and choose Q so that m_Q is odd (if all m_P are even, $\rho(C) = 0$). Then $C_0^{[Q]}$ has dimension $k - 1$. Let $\hat{C}$, the contracted code of $C_0^{[Q]}$, have parameters $[\nu, \kappa]$. Without loss of generality we may assume the generator matrix for C has the form shown in Fig. 1. The parities of the multiplicities m_Q, m_R, m_S, $\cdots$ are indicated below the matrix. QRS is a typical line with m_R, m_S odd; QTU is a typical line with m_T odd and m_U even; and QVW is a typical line with m_V, m_W even. There are η such pairs R, S, and ν such pairs T, U. x, y, z denote column vectors of length $k - 1$.

Then $\tilde{k}$ is the rank of the subspace of $\mathbf{F}^k$ spanned by the columns of types Q, R, S and T, while κ is the rank of the subspace of $\mathbf{F}^{k-1}$ spanned by the vectors y. The difference in rank $\Delta = \tilde{k} - \kappa$ therefore satisfies $1 \leqq \Delta \leqq 1 + \eta$. By the pivoting bound, $\rho(C) \leqq \eta + \rho(C_0^{[Q]})$, and by the induction hypothesis $\rho(C) \leqq \eta + T[\nu, \kappa]$. The desired result will

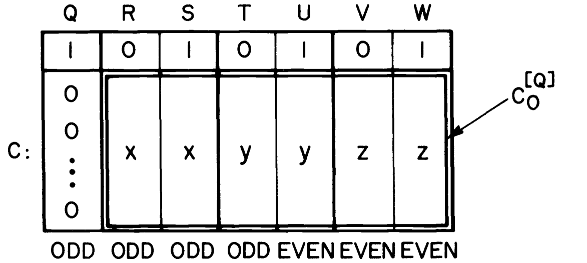


FIG. 1. *Pivoting* (*see proof of Theorem* 6).

follow if we show that

$$\eta + T[\nu, \kappa] \leqq T[\tilde{n}, \tilde{k}]. \tag{32}$$

We also know (by counting the odd columns in Fig. 1) that $\tilde{n} = 1 + 2\eta + \nu$. By (23), (24) we can find a code A with parameters $[2\eta + 1, \Delta]R = \eta$ (using (23) if $1 \leqq \Delta \leqq \eta$, and (24) if $\Delta = \eta + 1$). Suppose B is a code with parameters $[\nu, \kappa]$ and covering radius $R = T[\nu, \kappa]$. Then $A \oplus B$ has length $1 + 2\eta + \nu = \tilde{n}$, dimension $\kappa + \Delta = \tilde{k}$, and covering radius $\eta + T[\nu, \kappa]$. This implies (32) and completes the proof.

By combining Theorem 8 of [10] with Theorem 6, we obtain our final bound.

THEOREM 7. *Given a code C, let the contracted code $\tilde{C}$ have parameters $[\tilde{n}, \tilde{k}]$. Then*

(a) $\rho(C) \leqq [\tilde{n}/2] - 1$ *if $\tilde{k} \leqq [\tilde{n}/2]$ and $\tilde{C}$ is not a simplex code*;

(b) $\rho(C) = [\tilde{n}/2]$ *if $\tilde{k} \leqq [\tilde{n}/2]$ and $\tilde{C}$ is a simplex code (this requires $\tilde{n} = 2^i - 1$ for some i)*;

(c) $\rho(C) \leqq \tilde{n} - \tilde{k}$ *if $\tilde{k} \geqq \lceil \tilde{n}/2 \rceil$.*

Remark. We conjecture, but have not been able to prove, that Theorem 6 can be replaced by $\rho(C) \leqq T^*[\tilde{n}, \tilde{k}]$, where $T^*[\tilde{n}, \tilde{k}]$ is the largest covering radius of any *projective* $[\tilde{n}, \tilde{k}]$ code.

5. Behavior of ρ as $n \to \infty$; structure of best covering codes. In this section we investigate the behavior of ρ as the length $n \to \infty$ (while the dimension k is held fixed). There are only a finite number of projective codes of a given dimension k (for their length cannot exceed $2^k - 1$). Previously we began with an arbitrary code C and considered the contracted code $\tilde{C}$. Now we reverse the process and begin with a projective code B, with parameters $[n_B, k]$, and consider an arbitrary $[n, k]R$ code C, with multiplicities m_i say, for which $\tilde{C} = B$.

Note that R and $\rho(C)$ are related via (2):

$$\rho(C) = R - \frac{n}{2} + \frac{n_B}{2}. \tag{33}$$

Therefore, if n_B and n are fixed, by minimizing $\rho(C)$ we minimize R.

If $\tilde{C} = B$ and B is stable (§ 3), then $\rho(C) = \rho(B)$; in general, however, $\rho(C) > \rho(B)$. We define[3]

$$\rho_\infty(B) = \max \{\rho(C): C \equiv B, C > B\}. \tag{34}$$

Note that in view of Theorem 1(b), adjoining pairs of identical columns from outside the subspace of $PG(2, k - 1)$ has no effect on ρ, and so there is no loss of generality in (34) in assuming that dim C = dim B. We also note that by (22), $\rho_\infty(B)$ is finite:

$$\rho_\infty(B) \leqq \left[\frac{n_B}{2}\right]. \tag{35}$$

From the pivoting bound (26),

$$\rho_\infty(B) \leqq \eta + \rho_\infty(\hat{B}). \tag{36}$$

[3] This definition of ρ_∞ differs from that in [10] in allowing even multiples of columns not present in B to be adjoined to B.

This gives a useful criterion for stability. If, for some choice of the pivot,

$$\rho(B)=\eta+\rho_\infty(\hat{B}), \tag{37}$$

then B is stable. For (17) and (37) assume that $\rho_\infty(B)=\rho(B)$. Also, if $\tilde{C}=B$ and (37) holds, C is tame and stable.

The problem of finding $\rho_\infty(B)$ can also be expressed as an integer programming problem:

$$\rho_\infty(B)=\max\{h_1+\cdots+h_{2^k-1}\}, \tag{38}$$

subject to the constraints

$$h_i\in\mathbf{Z}\quad \text{for } i=1,\cdots,2^k-1, \tag{39}$$

$$\sum_{i=1}^{2^k-1} h_ic_i\leqq\frac{1}{2}\sum_{i=1}^{2^k-1}\pi_ic_i\quad \text{for } c=(c_1,\cdots,c_{2^k-1})\in S_k, \tag{40}$$

where the c_i (=0 or 1) in (40) are interpreted as real numbers, and $\pi_i=0$ if m_i is even, $\pi_i=1$ if m_i is odd (cf. [10, Thm. 1]).

If C is such that $\tilde{C}=B$, dim C = dim B, and $\rho(C)=\rho_\infty(B)$, we say that C has been obtained by *saturating* B. Obviously a saturated code is stable.

It is sometimes useful to know how long it would take to saturate B, if pairs of identical columns were added to B so as to drive ρ up to $\rho_\infty(B)$ as quickly as possible. The next theorem gives an upper bound on the answer, and also on the values of the m_i and h_i that are required.

THEOREM 8. *Let B be a projective code of dimension k.* (a) *There is an $[n, k]$ code C with multiplicities $m_i\leqq 2^k$ and length*

$$n\leqq 2^k(2^k-1), \tag{41}$$

satisfying $\tilde{C}=B$ and $\rho(C)=\rho_\infty(B)$. (b) *Consider any $[n, k]$ code C with $\tilde{C}=B$ and $\rho(C)=\rho_\infty(B)$. Let x be a deep hole in C for which* 0 *is a closest codeword. Then the heights h_i of the blocks of x satisfy*

$$|h_i|\leqq 2^{k-1},\qquad i=1,\cdots,2^k-1. \tag{42}$$

Proof. We prove (b) first. We know that

$$h_1+\cdots+h_{2^k-1}=\rho_\infty(B), \tag{43}$$

so we shall maximize h_i (and later $-h_i$) subject to the constraints (39), (40), (43). Inequality (40) implies that

$$\sum_{i=1}^{2^k-1} h_ic_i\leqq 2^{k-2}\quad \text{for } c\in S_k, \tag{44}$$

since all the nonzero codewords in S_k have weight 2^{k-1}, and $\pi_i\leqq 1$. If the inequalities (44) corresponding to those codewords $c=(c_1,\cdots,c_{2^k-1})\in S_k$ with $c_1=1$ are summed we obtain $2^{k-1}h_1+2^{k-2}(h_2+\cdots+h_{2^k-1})\leqq 2^{2k-3}$, which by (43) becomes $h_1\leqq 2^{k-1}-\rho_\infty(B)\leqq 2^{k-1}$. Similarly, by using the codewords with $c_1=0$, we obtain $-h_1\leqq 2^{k-1}$, so $|h_1|\leqq 2^{k-1}$. The same argument applies to the other h_i. (a) Knowing the possible range of the individual h_i we can work out how large the multiplicities m_i must be. From (9), $m_i=2^k$ is enough to permit $-2^{k-1}\leqq h_i\leqq 2^{k-1}$. Therefore there is a code C of length $n=\sum m_i\leqq 2^k(2^k-1)$ with $\rho(C)=\rho_\infty(B)$.

Theorem 8 implies that we can add the extra inequalities

$$(45) \qquad |h_i| \leqq 2^{k-1}, \qquad i=1, \cdots, 2^k-1,$$

to the integer program (38)–(40) defining $\rho_\infty(B)$. Stronger bounds than (45) can often be obtained in particular cases by detailed examination of (40).

The quantity $\rho_\infty(B)$ tells us the *maximal* value of $\rho(C)$ over all codes C with $\tilde{C} = B$. We now investigate the *minimal* value. Let

$$(46) \qquad \rho_n^*(B) = \min \{\rho(C)\text{: } C \text{ is an } [n,k] \text{ code with } C \equiv B, C \geqq B\}.$$

In view of (33), this also minimizes the covering radius of C over all $[n, k]$ codes with $\tilde{C} = B$. From $\rho_n^*(B) \leqq \rho_\infty(B)$ and (15), for sufficiently large n $\rho_n^*(B)$ is independent of n, and will be denoted by $\rho_*(B)$. Then

$$(47) \qquad \rho(B) \leqq \rho_*(B) \leqq \rho_\infty(B).$$

THEOREM 9. *Let B be an $[n_B, k]$ projective code. For sufficiently large n of the same parity as n_B there is an $[n, k]$ normal code C with $C \equiv B$, $C \geqq B$, and $\rho(C) = \rho_*(B)$, which is obtained by adjoining $n - n_B$ copies of a single column to B.*

Proof. Let us arrange the coordinates of S_k so that the coordinates of B appear first. The set of vectors $(h_1, \cdots, h_{2^k-1})$ satisfying (39), (40) (with $\pi_i = 1$ for $1 \leqq i \leqq n_B$, $\pi_i = 0$ otherwise) and (45) is a finite set H (depending on B but not on n).

For fixed large n with $n \equiv n_B (\text{mod } 2)$, we consider all $[n, k]$ codes C for which $\tilde{C} = B$. Such a code C is defined by its multiplicities $m_1, \cdots, m_{2^k-1}$, where

$$(48) \qquad \begin{aligned} &m_1, \cdots, m_{n_B} \quad \text{are odd,} \\ &m_{n_B+1}, \cdots, m_{2^k-1} \quad \text{are even,} \quad \text{and} \\ &\sum_{i=1}^{2^k-1} m_i = n. \end{aligned}$$

Then $\rho(C) = \rho^{(k)}(m_1, \cdots, m_{2^k-1})$ is equal to the maximum of $h_1 + \cdots + h_{2^k-1}$ subject to (39), (40) and

$$(49) \qquad -\left[\frac{m_i}{2}\right] \leqq h_i \leqq \left[\frac{m_i}{2}\right] \quad \text{for } i=1, \cdots, 2^k-1$$

(cf. [10, Thm. 1]). Also

$$\rho_n^*(B) = \min \{\rho^{(k)}(m_1, \cdots, m_{2^k-1})\text{: the } m_i \text{ satisfy (48)}\}.$$

Let us define

$$f(i) = \rho^{(k)}(1, \cdots, 1, n-n_B+1, 1, \cdots, 1, 0, \cdots, 0),$$

for $i = 1, \cdots, n_B$, where $n - n_B + 1$ is in position i and there are $n_B - 1$ ones, and

$$f(i) = \rho^{(k)}(1, \cdots, 1, 0, \cdots, 0, n-n_B, 0, \cdots, 0)$$

for $i = n_B + 1, \cdots, 2^k - 1$, where $n - n_B$ is in position i and there are n_B ones. Since H is a finite set, for n sufficiently large $f(i)$ is independent of n and is equal to either

$$\rho^{(k)}(1, \cdots, 1, a_i, 1, \cdots, 1, 0, \cdots, 0)$$

for $i = 1, \cdots, n_B$, or

$$\rho^{(k)}(1, \cdots, 1, 0, \cdots, 0, a_i, 0, \cdots, 0)$$

for $i = n_B + 1, \cdots, 2^k - 1$, where $a_1, \cdots, a_{2^k-1}$ are constants (depending only on B).

Finally, let

$$M = \min\{f(1), \cdots, f(2^k-1)\}$$

and choose an r such that $f(r) = M$.

After these preliminaries we come to the heart of the proof. Suppose n is large (specifically, we need $n > 2^k \max\{a_i\}$), and $\rho_n^*(B) = \rho^{(k)}(m_1, \cdots, m_{2^k-1})$, where the m_i satisfy (48). At least one of the m_i must be large, say m_j (and in particular $m_j > a_j$). By the monotonicity property (15),

$$\begin{aligned}\rho_n^*(B) &\geqq \begin{cases}\rho^{(k)}(1, \cdots, m_j, \cdots, 1, 0, \cdots, 0) & (j \leqq n_B)\\ \rho^{(k)}(1, \cdots, 1, 0, \cdots, m_j, \cdots, 0) & (j > n_B)\end{cases}\\ &= \begin{cases}\rho^{(k)}(1, \cdots, a_j, \cdots, 1, 0, \cdots, 0) & (j \leqq n_B)\\ \rho^{(k)}(1, \cdots, 1, 0, \cdots, a_j, \cdots, 0) & (j > n_B)\end{cases}\\ &\geqq M.\end{aligned}$$

In fact $\rho_n^*(B) = M$, since if C has multiplicities $m_i = 1$ for $i \leqq n_B$ and $m_i = 0$ for $i > n_B$, except for $m_r = n - n_B + 1$ (if $r \leqq n_B$) or $n - n_B$ (if $r > n_B$), then $\rho(C) = M$. Therefore $\rho_n^*(B) = \rho_*(B) = M$.

Since $\rho(C) = \rho_*(B)$, we can adjoin $2l$ copies of the special column to C without increasing ρ, for $l = 0, 1, \cdots$. Therefore C is normal by Lemma 5. This completes the proof of Theorem 9.

Remark. The single column mentioned in the theorem need not be a column of B.

Example. As an illustration of the preceding notions we analyze the $[11, 5]R = 3$ code C'_{11} having generator matrix

$$C'_{11}: \quad \begin{array}{|ccccccccccc|} \multicolumn{1}{c}{1} & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & \multicolumn{1}{c}{11} \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ \hline \end{array}\ , \tag{50}$$

which is an interesting code for several reasons (see § 6). It has weight distribution $0^1\,4^{10}\,6^{16}\,8^5$, and its automorphism group has order 1920, with two orbits on coordinates, $\{1\}$ and $\{2, \cdots, 11\}$. Suppose C is an $[n, 5]$ code with $\tilde{C} = C'_{11}$, having odd multiplicities $m_1, \cdots, m_{11}$ on the columns of (50), and even multiplicities on the remaining 20 nonzero columns of length 5. From (17), $\rho(C) \geqq \rho(C'_{11}) = 3$.

If we pivot on coordinate $Q = 1$, we see that $\eta = 5$ (since columns 2 and 7, 3 and 8, etc. combine), $C_0^{[1]}$ has even multiplicities; i.e., $\hat{C} = 0$, and (26) states that $\rho(C) \leqq 5$. On the other hand, if we pivot at $Q = 2$ we find that $\eta = 1$, $\hat{C}$ is the [8, 4] extended Hamming code, for which $\rho = 2$ or 3 depending on the multiplicities m'_P [10, Thm. 16], and (26) yields $\rho(C) \leqq 4$. We conclude that

$$3 \leqq \rho(C) \leqq 4. \tag{51}$$

The following theorem determines which of these two possibilities occurs.

THEOREM 10. *Let C be any code for which the contracted code is the* $[11, 5]R = 3$ *code C'_{11} defined by* (50). *If column* 1 *of* (50) *has multiplicity* $\geqq 3$ *or any of the* 20 *nonzero* *columns of length* 5 *not in* (50) *have positive multiplicity, then* $\rho(C) = 4$; *otherwise* $\rho(C) = 3$. *Thus* $\rho_*(C'_{11}) = 3$, $\rho_\infty(C'_{11}) = 4$.

Proof. To check the first assertion we verify (by computer) that if two copies of column 1 or of any of the 20 missing columns are adjoined to (50), ρ increases to 4, and then we use monotonicity (15). On the other hand, by integer programming we find that $\rho(C) = 3$ in all other cases.

For example, by assigning multiplicities 1, 1, 1, 1, m, 1, $\cdots$, 1 (m odd) to the columns of (50) we obtain an infinite sequence of $[n = m + 10, 5]$ codes with $\rho = 3$ and covering radius $R = 3 + [m/2] = (n - 5)/2$, for odd $n \geqq 11$. Figure 2 shows the case of length 23. These codes are optimal coverings, for it is proved in Theorem 22 of [3] that $t[n, 5] = [(n - 5)/2]$ ($n \neq 6$). (They are not unique, however; there are many codes that achieve this bound).

Remark. Definitions (34), (46) and Theorem 9 still apply if B is not projective (although the proof of Theorem 9 must be modified slightly).

THEOREM 11. *If C is normal, then* $\rho_*(C) = \rho(C)$.

Proof. Suppose column i of C is acceptable. By adjoining $2l$ copies of column i to C we obtain a code D with $\rho(D) = \rho(C)$ (see Theorem 2). Therefore $\rho_*(C) = \rho(C)$.

Remark. Similarly, if C is a code of length n with the property that, for all $i = 1, \cdots, n$, adjoining two copies of column i to C increases ρ, then C is abnormal.

Theorem 9 also provides information about the best possible covering codes.

THEOREM 12. *For fixed k, and all sufficiently large n, then* (a)

$$t[n,k] = \frac{n}{2} + \min_B \left\{ \rho_*(B) - \frac{n_B}{2} \right\}, \tag{52}$$

where B ranges over all projective codes of dimension k or $k - 1$ (a finite set), and n_B is the length of B; (b) *there is a normal $[n, k]R$ code C with $R = t[n, k]$ in which all columns have multiplicity* 1 *except for one column which has large multiplicity*; *and* (c)

$$t[n+2,k] = t[n,k] + 1. \tag{53}$$

Proof. Suppose $n \gg k$, and let $\mathscr{C}[n, k]$ denote the set of all $[n, k]$ codes with covering radius $R = t[n, k]$. Choose any $C \in \mathscr{C}[n, k]$ and let $B = \tilde{C}$ be an $[n_B, k_B]$ code. Then $\rho(C) \geqq \rho_*(B)$. By Theorem 9 there is an $[n - 2(k - k_B), k_B]$ normal code D with $\tilde{D} = B$ and $\rho(D) = \rho_*(B)$. Then $C' = D \oplus T_2^{k-k_B}$ is an $[n, k]$ normal code with contracted code B, and $\rho(C') = \rho_*(B)$. Thus $C' \in \mathscr{C}[n, k]$ and $\rho(C) = \rho(C') = \rho_*(B)$. From (2), $CR(C) = \frac{1}{2}n - \frac{1}{2}n_B + \rho_*(B)$, and therefore

$$t[n,k] = \frac{n}{2} + \min_B \left\{ \rho_*(B) - \frac{n_B}{2} \right\},$$

where B ranges over all projective codes of dimension $k_B \leqq k$.

We next show that in fact $k_B = k$ or $k - 1$, and that there is a normal code $C'' \in \mathscr{C}[n, k]$ in which all but one of the columns has multiplicity 1.

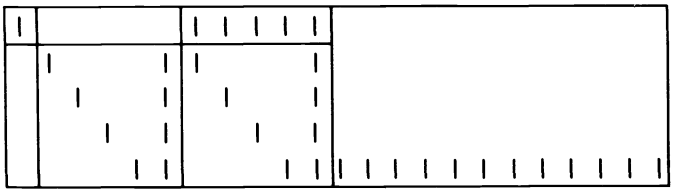


FIG. 2. *A* [23, 5]*R* = 9 *optimal covering code obtained from* (50). *Blank entries indicate zeros.*

Case (i): $k \equiv k_B \pmod 2$. By Theorem 9 there is an $[n-(k-k_B), k_B]$ normal code A with $\tilde{A} = B$ and $\rho(A) = \rho_*(B)$, obtained by adjoining $(n-n_B)-(k-k_B)$ copies (an even number) of a single column β (say) to B. Then $C'' = A \oplus \mathbf{F}^{k-k_B}$ is an $[n, k]$ normal code with $\rho(C'') = \rho_*(B)$, and has $n_B + k - k_B$ distinct columns with odd multiplicity. From (2),

$$CR(C'') = \frac{n}{2} - \frac{n_B + k - k_B}{2} + \rho_*(B)$$

which is less than $CR(C)$ unless $k = k_B$. Therefore $k = k_B$, and $C'' \in \mathscr{C}[n, k]$ has the desired multiplicities.

Case (ii): $k \not\equiv k_B \pmod 2$. By Theorem 9, for sufficiently large N_0 there is an $[N_0, k_B]$ normal code A_0 with $\tilde{A}_0 = B$ and $\rho(A_0) = \rho_*(B)$, obtained by adjoining $N_0 - n_B$ copies of a single column β to B. Let A_i be obtained by adjoining i further copies of β to A_0. We know $CR(A_2) = CR(A_0) + 1$, so

$$\begin{aligned} &\text{either} \quad CR(A_1) = CR(A_0), \qquad CR(A_2) = CR(A_1) + 1, \\ &\text{or} \qquad CR(A_1) = CR(A_0) + 1, \quad CR(A_2) = CR(A_1). \end{aligned}$$

In the first case we call A_0 *late* and in the second case we call it *early*. Whether A_0 is late or early depends on the solution to a certain integer programming problem. Therefore, in the sequence $A_0, A_2, A_4, \cdots$, from a certain point on either all A_{2i} are early or all are late, and similarly in the sequence $A_1, A_3, A_5, \cdots$. Thus for sufficiently large i, A_{2i+1} satisfies the hypothesis of Lemma 5 and is normal. In particular, by taking

$$i = \tfrac{1}{2}\{n - 1 - N_0 - (k - k_B)\},$$

we obtain normal codes A_{2i} and A_{2i+1} of dimension k_B and lengths $n-1-(k-k_B)$ and $n-(k-k_B)$, respectively, with $\rho(A_{2i}) = \rho_*(B)$,

$$CR(A_{2i}) = \frac{n-1-(k-k_B)}{2} - \frac{n_B}{2} + \rho_*(B),$$

$$CR(A_{2i+1}) \leqq CR(A_{2i}) + 1.$$

Finally, let $C'' = A_{2i+1} \oplus \mathbf{F}^{k-k_B}$. Then

$$CR(C'') = CR(A_{2i+1}) \leqq \frac{n}{2} - \frac{n_B}{2} + \rho_*(B),$$

since $k \geqq k_B + 1$, with equality only if $k = k_B + 1$ and A_{2i} is early. Therefore $k = k_B + 1$, and $C'' \in \mathscr{C}[n, k]$ has the desired multiplicities. This completes the proof of (a) and (b). To prove (c) we observe that the best choice for B in (52) is independent of n, and when n increases to $n+2$, the right-hand side of (52) increases by 1.

Remarks. (1) This theorem establishes Conjectures A and D of [2] for sufficiently large n. (2) If the optimal code C is obtained by adjoining $2l$ columns to B, then C has length $n = n_B + 2l$ and covering radius $R = \rho_*(B) + l$. We can write this as

$$\frac{n}{2} - R = \frac{n_B}{2} - \rho_*(B) \tag{54}$$

or, if B is normal, as

$$\frac{n}{2} - CR(C) = \frac{n_B}{2} - CR(B). \tag{55}$$

Thus the best seed codes B are normal codes for which the parameter

$$\delta = \frac{n_B}{2} - CR(B) \tag{56}$$

is as large as possible. (3) A possible interpretation of Theorems 9 and 12 is the following. An optimal covering code has the property that the codewords are constructed so as to be not too far from an arbitrary n-tuple x. This is a difficult task for $n \gg k$, since we are using only 2^k vectors to cover 2^n vectors. We may say informally that codes with the structure described in Theorem 12(b) do this by matching x very carefully on a small number (n_B) of coordinates, and just using an average on the rest (see Fig. 2). (4) This special structure also greatly simplifies the process of finding the closest codeword to a given x.

REFERENCES

[1] G. D. Cohen, M. G. Karpovsky, H. F. Mattson, Jr. and J. R. Schatz, *Covering radius—survey and recent results*, IEEE Trans. Inform. Theory, IT-31 (1985), pp. 328–343.

[2] G. D. Cohen, A. C. Lobstein and N. J. A. Sloane, *Further results on the covering radius of codes*, IEEE Trans. Inform. Theory, IT-32 (1986), pp. 680–694.

[3] R. L. Graham and N. J. A. Sloane, *On the covering radius of codes*, IEEE Trans. Inform. Theory, IT-31 (1985), pp. 385–401.

[4] H. J. Helgert and R. D. Stinaff, *Minimum-distance bounds for binary linear codes*, IEEE Trans. Inform. Theory, IT-19 (1973), pp. 344–356.

[5] J. S. Leon, *Computing automorphism groups of error correcting codes*, IEEE Trans. Inform. Theory, IT-28 (1982), pp. 496–511.

[6] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[7] H. F. Mattson, Jr., *An Improved Upper Bound on Covering Radius*, Lecture Notes Computer Science 228, 1986, pp. 90–106.

[8] D. Slepian, *Some further theory of group codes*, Bell System Tech. J., 39 (1960), pp. 1219–1252.

[9] N. J. A. Sloane, *Unsolved problems related to the covering radius of codes*, in Proc. Specific Problems in Communication and Computation (SPOC-85), Springer-Verlag, New York, 1987, to appear.

[10] ———, *A new approach to the covering radius of codes*, J. Combin. Theory Ser. A, 42 (1986), pp. 61–86.

[11] T. Verhoeff, *Updating a table of bounds on the minimum distance of binary linear codes*, Math. Dept., Eindhoven Univ. Tech., Report 85-WSK-01, Eindhoven, the Netherlands, January 1985; IEEE Trans. Inform. Theory, to appear.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

008

ON THE COVERING RADIUS PROBLEM FOR CODES II. CODES OF LOW DIMENSION; NORMAL AND ABNORMAL CODES*

KAREN E. KILBY† AND N. J. A. SLOANE‡

Abstract. In this two-part paper we introduce the notion of a stable code and give a new upper bound on the normalized covering radius of a code. The main results are that, for fixed k and large n, the minimal covering radius $t[n, k]$ is realized by a normal code in which all but one of the columns have multiplicity 1; hence $t[n + 2, k] = t[n, k] + 1$ for sufficiently large n. We also show that codes with $n \leqq 14$, $k \leqq 5$ or $d_{\min} \leqq 5$ are normal, and we determine the covering radius of all proper codes of dimension $k \leqq 5$. Examples of abnormal nonlinear codes are given. In Part I [this Journal, 8 (1987), pp. 604–618] we investigated the general theory of normalized covering radius, while in Part II we study codes of dimension $k \leqq 5$, and normal and abnormal codes.

Key words. binary codes, covering radius

AMS(MOS) subject classifications. 05B, 94B

6. Codes of dimension $k \leqq 5$.

The inequivalent projective $[n, k]$ codes were enumerated by Slepian [8]. The numbers of these codes for $k \leqq 5$ are shown in Table 1, as well as the numbers of inequivalent, projective, *indecomposable* $[n, k]$ codes. The codes of dimension $k \leqq 4$ were studied in detail in [10], where it was shown that, with four exceptions, all such codes are tame and stable. The exceptions are described in Theorem 16 of [10], and are also listed below. We have made a similar investigation of all the proper codes of dimension 5. It would take too much space to list all the codes, and we just give a summary. The main result is the following.

THEOREM 13. *For each projective code B of dimension* 5,

$$CR(B) \leqq \rho_\infty(B) \leqq CR(B) + 1. \tag{57}$$

Equivalently, for any code C for which the contracted code $\tilde{C}$ has dimension 5,

$$CR(\tilde{C}) \leqq \rho(C) \leqq CR(\tilde{C}) + 1. \tag{58}$$

Every code of dimension $\leqq 5$ is normal, and in fact every code C for which the contracted code $\tilde{C}$ has dimension $\leqq 5$ is normal.

Proof. The first assertion was established by direct calculation using a computer, the pivoting bound (26) of Part I, and the results of [10]. The normality of these codes then follows from Theorem 9 of [10] (see the remarks in § 4 of Part I), and the final assertion is a consequence of Theorem 1(b) of Part I.

Remark. While proving Theorem 13 we reconfirmed the assertions required for the proof of Theorem 22 of [3] that there are no proper codes with parameters [11, 5] 2, [13, 5] 3, $\cdots$, [23, 5] 8 or [25, 5] 9.

Table 2(a) gives upper and lower bounds on $\rho(C)$ for any code C for which $\tilde{C}$ is an $[n, k]$ code with $k \leqq 5$. Table 2(b) gives $t[n, k]$ for $k \leqq 5$ (a much more extensive table appears in [3]). For large n (indicated by an asterisk) $t[n, k]$ cannot be realized by a projective code.

The four wild codes of dimension 4 are the [7, 4] 1 Hamming code H_7, the [8, 4] 2 extended Hamming code, the [11, 4] 3 code obtained by omitting the columns of an

* Received by the editors February 6, 1986; accepted for publication (in revised form) March 24, 1987.

† Morse College, Yale University, New Haven, Connecticut 06520. The work of this author was supported in part by Bell Laboratories.

‡ Mathematical Sciences Research Center, AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

TABLE 1

Number of inequivalent projective [*n*, *k*] *codes.* (*In parentheses, the number of inequivalent, projective, indecomposable* [*n*, *k*] *codes.*)

$k \backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1 (1)														
2		1 (0)	1 (1)												
3			1 (0)	2 (1)	1 (1)	1 (1)	1 (1)								
4				1 (0)	3 (1)	4 (2)	5 (4)	6 (5)	5 (5)	4 (4)	3 (3)	2 (2)	1 (1)	1 (1)	1 (1)
5					1 (0)	4 (1)	8 (3)	15 (9)	29 (22)	46 (40)	64 (60)	89 (86)	112 (110)	128 (127)	144 (143)

$k \backslash n$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
5	145 (144)	129 (129)	113 (113)	91 (91)	67 (67)	50 (50)	34 (34)	21 (21)	14 (14)	9 (9)	5 (5)	3 (3)	2 (2)	1 (1)	1 (1)	1 (1)

$E_3 \oplus \mathbf{F}$ from S_4, and the [12, 4] 4 code obtained by omitting an E_3 from S_4. For all four codes, $\rho_\infty = R + 1$. (See Theorem 16 of [10].)

Projective codes of dimension 5. In the remainder of this section we briefly describe the [n, 5] projective codes for each length n, and give one or two of the best covering codes. The examples given were selected by applying (in order) the following criteria: (1) minimize R, (2) choose tame rather than wild codes, (3) maximize the order g of the automorphism group, and (4) maximize $d_{\min}$. Generator matrices for some of these codes are displayed in Fig. 3. Although many of these codes can be obtained from later codes by deleting appropriate coordinates, there is no single sequence of nested codes that includes all our best examples.

For each code B in the following list we give upper and lower bounds (differing by at most 1) on $\rho(C)$ for any code C for which $\tilde{C} = B$. All tame codes of dimension 5 satisfy (37) of Part I and are therefore stable. However we have not been able to check all the wild codes (not even all the codes on the following list) to see if they are unstable and the upper bound on $\rho(C)$ is actually attained.

$n = 5$. One code: $\mathbf{F}^5$, $\rho = 0$, tame, $g = 5!$, $d_{\min} = 1$.

$n = 6$. All four are tame, with $R = 1$. Example: E_6, $\rho = 1$, tame, $g = 6!$, $d_{\min} = 2$.

$n = 7$. All are tame, six have $R = 1$, two have $R = 2$. Example: C_7 (Fig. 3), $\rho = 1$, tame, $g = 72$, $d_{\min} = 2$.

$n = 8$. Two wild codes, with $\rho = 1$ or 2, thirteen tame, with $R = 2$. Example: $H_7 \oplus \mathbf{F}$, $\rho = 1$ or 2, wild, $g = 168$, $d_{\min} = 1$. (This code is unstable. It can be shown (compare Theorem 10 of Part I) that the precise value of ρ is as follows. Consider any code C for which $\tilde{C} = H_7 \oplus \mathbf{F}$. The columns of H_7 span a $PG(2, 3)$, in which there is a unique further point X that is neither a column nor the sum of two columns of H_7. If

TABLE 2

(a) *Upper and lower bounds on normalized covering radius $\rho(C)$, given that $\tilde{C}$ is an $[n, k]$ code.* (b) *$t[n, k]$, the smallest possible covering radius R for any $[n, k]$ code. An asterisk indicates that this value of R cannot be realized by a projective code.*

(a)

$k\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0														
2		0	1												
3			0	1	1	2	3								
4				0	1	1-2	1-2	2-3	3	3-4	3-4	4	5	6	7
5					0	1	1-2	1-2	2-3	2-4	3-4	3-5	4-5	4-5	5-6

$k\backslash n$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
5	5-7	6-7	6-7	7-8	7-9	8-9	8-9	9-10	9-11	10-11	10-11	11-12	12-13	13	14	15

(b)

$k\backslash n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3			0	1	1	2	2*	3*	3*	4*	4*	5*	5*	6*	6*
4				0	1	1	1	2	2*	3	3	4	4*	5*	5*
5					0	1	1	1	2	2	3	3	4	4	5

$k\backslash n$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
3	7*	7*	8*	8*	9*	9*	10*	10*	11*	11*	12*	12*	13*	13*	14*	14*
4	6*	6*	7*	7*	8*	8*	9*	9*	10*	10*	11*	11*	12*	12*	13*	13*
5	5	6	6	7	7	8	8	9	9	10	10	11	11*	12*	12*	13*

any of the seven points $PG(2, 3)\backslash\{X \cup H_7\}$ has positive even multiplicity in C, $\rho(C) = 2$; otherwise $\rho(C) = 1$.)

$n = 9$. Twenty-three have $R = 2$ (2 are wild), six have $R = 3$ (all tame). Example: C_9, $\rho = 2$, tame, $g = 384$, $d_{\min} = 2$.

$n = 10$. Seven have $R = 2$ (all wild), 38 have $R = 3$ (all tame), one has $R = 4$ (tame). Example: C_{10}, $\rho = 2$ or 3, wild, unstable ($\rho = 3$ if any column not in C_{10} has positive even multiplicity, otherwise $\rho = 2$), $g = 1920$, $d_{\min} = 4$.

$n = 11$. Fifty-five have $R = 3$ (2 are wild), nine have $R = 4$ (all tame). Example: C_{11}, $\rho = 3$, tame, $g = 120$, $d_{\min} = 4$.

$n = 12$. Twenty have $R = 3$ (all wild), 68 have $R = 4$ (all tame), one has $R = 5$ (tame). Example: C_{12} (Fig. 3, or omit first, second and last two columns of C_{16}), $\rho = 3$ or 4, wild, unstable, $g = 192$, $d_{\min} = 4$.

$n = 13$. One hundred four have $R = 4$ (6 are wild), eight have $R = 5$ (all tame). Example: C_{13} (omit first, second and last columns of C_{16}), $\rho = 4$, tame, $g = 576$, $d_{\min} = 5$.

$n = 14$. Forty-four have $R = 4$ (all wild), 88 have $R = 5$ (all tame). Example: C_{14} (omit first and last columns of C_{16}), $\rho = 4$ or 5, wild, unstable, $g = 2688$, $d_{\min} = 6$.

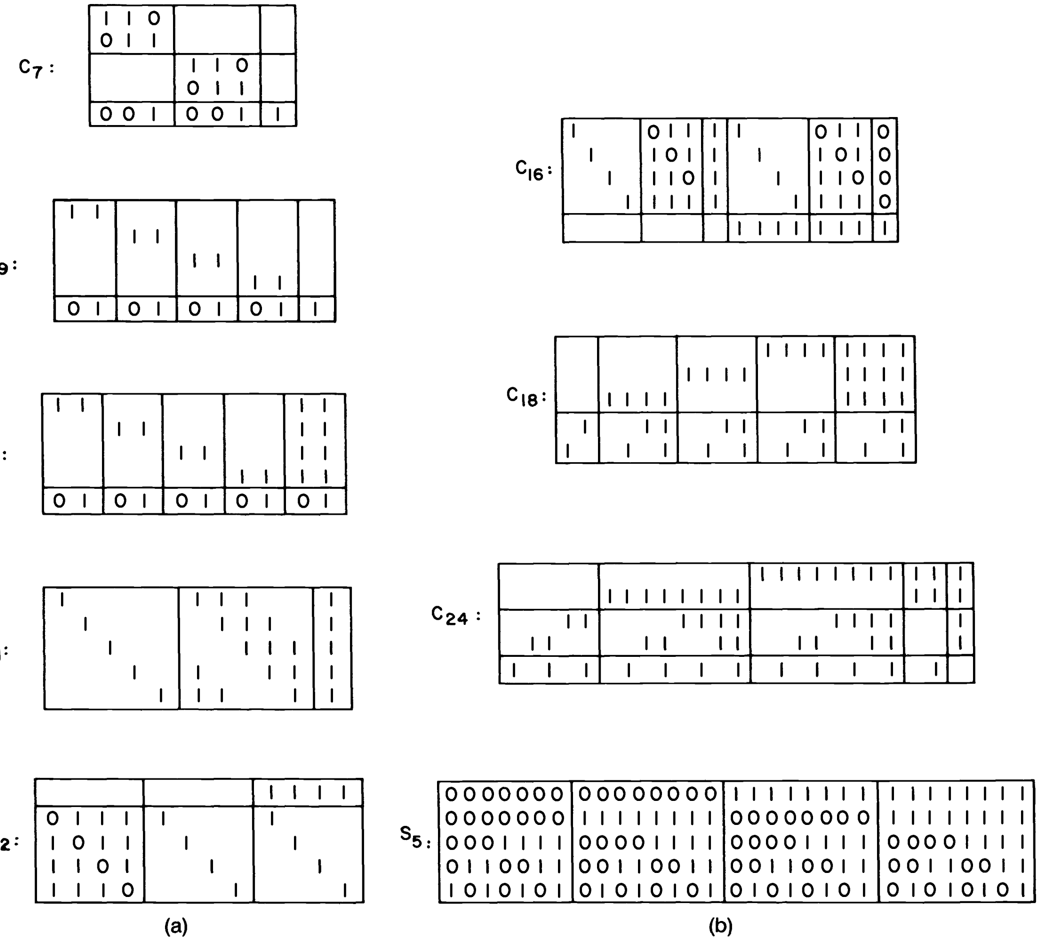


FIG. 3. *Selected best (proper) covering codes of dimension* 5. *Blank entries indicate zeros.*

$n = 15$. One hundred twenty-nine have $R = 5$ (9 are wild), 15 have $R = 6$ (all tame). Example: C_{15} (omit first column of C_{16}), $\rho = 5$, tame, $g = 192$, $d_{\min} = 6$. (There is an unstable code with the same R, and higher g and $d_{\min}$, namely the shortened first-order Reed–Muller code C'_{15}: $\rho = 5$ or 6, wild, unstable, $g = 20160$, $d_{\min} = 7$. Unstable codes with a larger group and $d_{\min}$ also occur at other lengths, and we mention some of these codes in parentheses.)

$n = 16$. Fifty-three have $R = 5$ (all wild), 89 have $R = 6$ (2 are wild), three have $R = 7$ (all tame). Example: C_{16} (Fig. 3, or omit first and last columns of C_{18}), $\rho = 5$ or 6, wild, unstable, $g = 1344$, $d_{\min} = 7$.

$n = 17$. One hundred nineteen have $R = 6$ (8 are wild), ten have $R = 7$ (all tame). Example: C_{17} (omit last column of C_{18}), $\rho = 6$, tame, $g = 192$, $d_{\min} = 7$. (Alternatively C'_{17}: omit first column of C_{18}, $\rho = 6$ or 7, wild, unstable, $g = 21504$, $d_{\min} = 8$.)

$n = 18$. Fifty-five have $R = 6$ (all wild), 58 have $R = 7$ (all tame). Example: C_{18} (Fig. 3, or omit columns 1–5 and last column of C_{24}), $\rho = 6$ or 7, wild, unstable, $g = 3072$, $d_{\min} = 8$.

$n = 19$. Eighty-six have $R = 7$ (8 are wild), five have $R = 8$ (all tame). Example: C_{19} (omit columns 1–5 of C_{24}), $\rho = 7$, tame, $g = 768$, $d_{\min} = 8$.

$n = 20$. Thirty-nine have $R = 7$ (all wild), 28 have $R = 9$ (1 is wild). Example: C_{20} (omit columns 1–4 of C_{24}), $\rho = 7$ or 8, wild, stability not known,[4] $g = 384$, $d_{\min} = 9$.

$n = 21$. Forty-seven have $R = 8$ (6 are wild), three have $R = 9$ (all tame). Example: C_{21} (omit columns 1–3 of C_{24}), $\rho = 8$, tame, $g = 384$, $d_{\min} = 9$. (Alternatively, the cyclic code C'_{21} with generator polynomial $x^{16} + x^{12} + x^{11} + x^8 + x^6 + x^4 + x^3 + x^2 + x + 1$, $\rho = 8$ or 9, wild, stability not known, $g = 1008$, $d_{\min} = 10$.)

$n = 22$. Nineteen have $R = 8$ (all wild), 15 have $R = 9$ (all tame). Example: C_{22} (omit columns 1 and 2 of C_{24}), $\rho = 8$ or 9, wild, $g = 768$, $d_{\min} = 10$.

$n = 23$. Nineteen have $R = 9$ (3 are wild), two have $R = 10$ (both tame). Example: C_{23} (omit columns 1–4, 6, 8, 9, 12 of S_5), $\rho = 9$, tame, $g = 128$, $d_{\min} = 10$. (Alternatively, C'_{23}: omit columns 1–8 of S_5, or column 1 of C_{24}, $\rho = 9$ or 10, wild, $g = 2688$, $d_{\min} = 11$.)

$n = 24$. Six have $R = 9$ (all wild), eight have $R = 10$ (1 wild). Example: C_{24} (Fig. 3, or omit columns 1–6, 8 of S_5), $\rho = 9$ or 10, wild, $g = 384$, $d_{\min} = 11$.

$n = 25$. Eight have $R = 10$ (2 are wild), one has $R = 11$ (tame). Example: C_{25} (omit columns 1, 2, 4, 7–9 of S_5), $\rho = 10$, tame, $g = 192$, $d_{\min} = 11$. (Alternatively C'_{25}: omit columns 1–6 of S_5, $\rho = 10$ or 11, wild, $g = 9216$, $d_{\min} = 12$.)

$n = 26$. Three have $R = 10$ (all wild), two have $R = 11$ (both tame). Example: C_{26} (omit first 5 columns of S_5), $\rho = 10$ or 11, wild, $g = 3072$, $d_{\min} = 12$.

$n = 27$. Three codes with $R = 11$ (two are wild). Example: C_{27} (omit $\mathbf{F}^4$ from S_5), $\rho = 11$, tame, $g = 384$, $d_{\min} = 12$.

$n = 28$. Two codes. C_{28} (omit $\mathbf{F}^3$ from S_5), $\rho = 12$, tame, $g = 2304$, $d_{\min} = 13$. (Alternatively, C'_{28}: omit E_3 from S_5, $\rho = 12$ or 13, wild, $g = 64512$, $d_{\min} = 14$.) For $n \geqq 28$ we can achieve a smaller R by using codes that are not projective (see Table 2(b)).

$n = 29$. One code: C_{29} (omit any 2 columns of S_5), $\rho = 13$, tame, $g = 21504$, $d_{\min} = 14$.

$n = 30$. One code: C_{30} (omit any column of S_5), $\rho = 14$, tame, $g = 322560$, $d_{\min} = 15$.

$n = 31$. One code: S_5 (Fig. 3), $\rho = 15$, tame, $g = 31 \cdot 30 \cdot 28 \cdot 24 \cdot 16 = 9999360$, $d_{\min} = 16$.

Codes with unacceptable coordinates. We have established certain facts about codes having unacceptable coordinates (the definition is given in § 1 of Part I), which we record here. All coordinates of all projective codes of dimension $k \leqq 4$ are acceptable. The first instance of a projective code with an unacceptable coordinate is the [10, 5] 2 code C'_{10} given in (4) of [3]. The second is the [11, 5] 3 code C'_{11} of (50) in Part I, obtained by adding an overall parity check to C'_{10}. The first coordinate of C'_{11} is unacceptable. These two are the only projective codes of dimension 5 and length $n \leqq 13$ with unacceptable coordinates. By adjoining an even number of copies of any acceptable coordinate to either C'_{10} or C'_{11}, we obtain $[n, 5]$ codes with unacceptable coordinates for all $n \geqq 10$. Hence by taking direct sums with $\mathbf{F}$ there are $[n, k]$ codes with unacceptable coordinates for all $n \geqq k + 5 \geqq 10$.

7. Codes with minimal distance 4 or 5 are normal.

It is shown in [2] that all codes with minimal distance $d_{\min} \leqq 3$ are normal, and any coordinate in the support of a codeword of minimal weight is acceptable. In this section we extend this result to codes with minimal distance 4 and 5.

We begin with two lemmas, the first of which is elementary. For any $x \in \mathbf{F}^n$, let x' denote the vector obtained by deleting the first coordinate of x. Suppose C is an $[n, k]R$ code, and let D be the $[n - 1, k']R'$ code $D = \{c' : c \in C\}$.

[4] Preliminary calculations suggest this code may be stable, with $\rho_\infty = 7$.

LEMMA 14. *For any* $x \in \mathbf{F}^n$, $d(x, C) \geqq d(x', C')$.

LEMMA 15. *Suppose C and D have the same covering radius R, and let $u \in \mathbf{F}^{n-1}$ be such that $d(u, D) = R$. Then there are codewords c_0, c_1 in C such that*

$$d(u, c'_0) = d(u, c'_1) = R,$$

the first coordinate of c_0 is 0, *and the first coordinate of c_1 is* 1.

Proof. Let $x = 0u$. Then $d(x, C) = R$ by Lemma 14, say $d(x, c_0) = R$. If $c_0 = 1c'$ then $d(c', u) = R - 1$, a contradiction, so $c_0 = 0c'$, with $d(c', u) = R$ as required. Similarly $x = 1u$ leads to c_1.

THEOREM 16. *A code with minimal distance* 4 *or* 5 *is normal, and all coordinates in the support of a codeword of minimal weight are acceptable.*

Proof. Suppose C is an $[n, k]R$ code containing the word $s = 11 \ldots 100 \ldots 0$ of weight $d_{\min} = 4$ or 5. We present the proof so that it applies simultaneously to both cases.

Let primes indicate that the first coordinate has been deleted, and set $D = \{c' : c \in C\}$. Since D has minimal distance 1 less than C, we may assume by induction that D is normal and the first coordinate of D is acceptable. The covering radius of D is either R or $R - 1$.

We shall show that the first coordinate of C is acceptable; i.e.,

$$d(x, C_0^{(1)}) + d(x, C_1^{(1)}) \leqq 2R + 1,$$

for all $x \in \mathbf{F}^n$, or equivalently that, for each x, there exists $c_0 \in C_0^{(1)}$, $c_1 \in C_1^{(1)}$ such that

$$d(x', c'_0) + d(x', c'_1) \leqq 2R. \tag{59}$$

Let $d(x', D) = R - m$, where $m = 0, 1, 2, \cdots$.

Case $m = 0$. Then (59) follows immediately from Lemma 15.

Case $m = 1$. Without loss of generality we suppose there is $c_0 = .\,0 \ldots \in C$ such that $d(x', c'_0) = R - 1$. If $x = .\,1 \ldots$ then $g = s + c_0$ satisfies $d(x', g') \leqq R + 1$ and (59) holds (with $c_1 = g$). So we may assume $x = .\,0 \ldots$, say $x = \alpha 0 \beta \gamma \delta \ldots$. Since D is normal there is $c_1 = .\,1 \ldots \in C$ such that $d(x', c'_1) \leqq R + 2$. We choose c_1 to minimize $d(x', c'_1)$. There are four subcases.

(a) $d(x', c'_1) = R + 2$. Let $y = \alpha 1 \beta \gamma \delta \ldots$ (differing from x in one coordinate), and let $c'_2 \in D$ be a closest codeword to y', with $d(y', c'_2) \leqq R$. Then $d(y', c'_2) = R$. (For suppose $d(y', c'_2) \leqq R - 1$. If $c_2 = .\,0 \ldots$, $d(x', c'_2) \leqq R - 2$, contradicting the definition of c_0, and if $c_2 = .\,1 \ldots$, $d(x', c'_2) \leqq R$, contradicting the definition of c_1.) By Lemma 15 there are $g_0 \in C_0^{(1)}$, $g_1 \in C_1^{(1)}$ with $d(y', g'_0) = d(y', g'_1) = R$. Also g_0 and g_1 have second digit 0, or else $d(x', g'_i) \leqq R + 1$, again contradicting the definition of c_1. Therefore $d(x', g'_0) = d(x', g'_1) = R - 1$, and (59) holds.

(b) $d(x', c'_1) = R + 1$. Define y and c_2 as in (a). Again $d(y', c'_2) = R$, and $g_0 \in C_0^{(1)}$, $g_1 \in C_1^{(1)}$ exist with $d(y', g'_0) = d(y', g'_1) = R$. If the second digit of either g_0 or g_1 is 0, $d(x', g'_0) + d(x', g'_1) \leqq 2R$, and (59) holds. Suppose then that $g_0 = 01 \ldots$, $g_1 = 11 \ldots$. Choose i so that g_i and c_0 differ in the first digit. Then $d(x', c'_0) + d(x', g'_i) \leqq (R - 1) + (R + 1)$, and again (59) holds.

(c) $d(x', c'_1) = R$. If $c_0 = \alpha 0 \ldots$ and $c_1 = \bar{\alpha} 1 \ldots$, or if $c_0 = \bar{\alpha} 0 \ldots$ and $c_1 = \alpha 1 \ldots$ then we may use c_0 and c_1 in (59). The difficult cases are when c_0 and c_1 begin with the same digit. We define $z = \alpha 0 \bar{\beta} \bar{\gamma} \delta \ldots$ (differing from x in two coordinates), let $d_0 = d(z, C) \leqq R$, and let U denote the set of codewords $u \in C$ with $d(z, u) = d_0$.

Case (c.1). Suppose $c_0 = \alpha 0 \ldots$, $c_1 = \alpha 1 \ldots$. If U contains a vector $u = \bar{\alpha} \ldots$ then $d(u, x) \leqq d(u, z) + 2 \leqq R + 2$, $d(u', x') \leqq R + 1$, and we use c_0 and u in (59). Otherwise all $u \in U$ begin with α, and $d(z', D') = d(z, D) = d_0$. If $d_0 = R$, then z is a deep hole in D and by Lemma 15 there is $u = \bar{\alpha} \ldots \in U$, a contradiction. Therefore $d_0 \leqq R - 1$. Choose some $u \in U$. If $u = \alpha . \beta\gamma \ldots$ then $d(x, u) \leqq d(z, u) - 2 \leqq R - 3$, a contradiction. Therefore $u = \alpha . \beta\bar{\gamma} \ldots, \alpha . \bar{\beta}\gamma \ldots$ or $\alpha . \bar{\beta}\bar{\gamma} \ldots$. In every case $g = s + u$ satisfies $d(x', g') \leqq R + 1$, and we may use c_0 and g in (59).

Case (c.2). Suppose $c_0 = \bar{\alpha}0 \ldots, c_1 = \bar{\alpha}1 \ldots$. The argument is similar to case (c.1) and is omitted.

(d) $d(x', c'_1) \leqq R - 1$. Let $g = s + c_1$. Then $d(x', g') \leqq R + 1$ and we may use c_1 and g in (59).

Case $m \geqq 2$. Choose $c \in C$ so that $d(x', c') \leqq R - 2$, say $c \in C_0^{(1)}$. Then $g = s + c \in C_1^{(1)}$ satisfies $d(x', g') \leqq R + 2$, and we use c and g in (59). This completes the proof.

8. Further conditions for a code to be normal. In this section we show that if $\tilde{C}$ has minimal distance 1 or 2 then C is normal. Theorem 18 summarizes the known conditions on the parameters of a code that imply normality.

THEOREM 17. *If the contracted code $\tilde{C}$ has $d_{\min} = 1$ or 2 then C is normal.*

Proof. If $\tilde{C}$ has $d_{\min} = 1$ then $\tilde{C}$ is a direct sum $\mathbf{F} \oplus \cdots$. Therefore $C = T_m \oplus \cdots$ for some repetition code T_m. Since T_m is normal, so is C by Theorem 1(b) of Part I.

Suppose $\tilde{C}$ has $d_{\min} = 2$, and there is a word of weight 2 with ones on the first and last coordinates. By Theorem 19 of [2], $\tilde{C}$ is normal, and the first and last coordinates are acceptable.

We first "blow up" all but the first and last coordinates of $\tilde{C}$, obtaining a code B (say) of length n_B, norm N_B, covering radius R_B, and $d_{\min} = 2$. Again by Theorem 19 of [2], B is normal and the first and last coordinates are acceptable.

Next we blow up the last coordinate of B, giving it odd multiplicity m_Q, and obtaining (by Theorem 2 of Part I) a normal code A (say) of length n_A, norm $N_A = N_B + m_Q - 1$, and covering radius $R_A = R_B + [m_Q/2]$. By Theorem 2 of Part I we know that the last m_Q coordinates of A are acceptable. The difficult part of the proof is to show that the *first* coordinate of A is acceptable. This is enough to prove the theorem, for then we can obtain C by applying the amalgamated direct sum construction again, blowing up the first coordinate of A, and deduce from Theorem 2 of Part I that C is normal.

Let $w'_2 \in B$ be the codeword with ones on the first and last coordinates of B, and let w_2 be the corresponding codeword of A. (Primes will indicate that the last $m_Q - 1$ coordinates have been deleted.) Take any $x \in \mathbf{F}^{n_A}$, and choose $b_i \in B_i^{(1)}$ so as to minimize $d(x', b_i)$ for $i = 0, 1$. Since the first coordinate of B is acceptable, $d(x', b_0) + d(x', b_1) \leqq N_B$. Let r, s be the last digits of b_0, b_1 respectively, and let $a_i \in A$ be obtained by repeating the last digit of b_i $m_Q - 1$ times. (α) If $r \neq s$ then the last digit of x' agrees with either r or s, and

$$d(x, a_0) + d(x, a_1) = d(x', b_0) + d(x', b_1) + m_Q - 1$$

$$\leqq N_B + m_Q - 1 = N_A,$$

and the first coordinate of A is acceptable. (β) If $r = s$, suppose the last digit of x' is $1 - r$. Then $y_0 = b_0 + w'_2$, $y_1 = b_1 + w'_2$ satisfy

$$d(x', y_0) + d(x', y_1) = d(x', b_0) + d(x', b_1) - 2,$$

contradicting the choice of b_0 and b_1. Therefore x' ends with r.

Let $b = b_0$ or b_1 be a closest codeword in B to x'. If the initial digits of x and b differ, let $c = b + w'_2 \in B$, so that b and c differ in their last digits, $d(x', b) = d(x', c)$, $d(x', b) + d(x', c) \leqq N_B$, and by ($\alpha$) the first coordinate of A is acceptable. Therefore we may assume x and b agree in their initial digits. We now have the situation shown in Fig. 4. If there are more than $\frac{1}{2}m_Q$ r's in the last m_Q digits of x, then $d(x, a_0) + d(x, a_1) \leqq d(x', b_0) + d(x', b_1) + m_Q - 1 \leqq N_A$. If there are fewer than $\frac{1}{2}m_Q$ r's we add w_2 to a_0 and a_1 and the same inequalities hold. This completes the proof.

THEOREM 18. *Let C be an $[n, k]R$ code, with minimal distance $d_{\min}$ and contracted code $\tilde{C}$. Any of the following conditions implies that C is normal:* (a) $n \leqq 14$, (b) $k \leqq 5$, (c) $R \leqq 2$, (d) $d_{\min} \leqq 5$, (e) *$\tilde{C}$ has dimension* $\leqq 5$, *or* (f) *$\tilde{C}$ has minimal distance* $\leqq 2$.

Proof. (b) and (e) follow from Theorem 13, (c) is Theorem 22 of [2], (d) follows from Theorem 24 of [2] and Theorem 16 above, and (f) is Theorem 17. (a) follows from (b), (d) and the tables of minimal distance [4], [11].

9. Abnormal nonlinear codes exist. At present it is not known if an abnormal *linear* code exists. Abnormal *nonlinear* codes were first constructed by Peter Frankl (personal communication), and with his permission we include his construction here.

In order for a code C to be abnormal, for each $i = 1, \cdots, n$ there must be a "bad" vector $x^{(i)}$ such that

(60) $$d(x^{(i)}, C_0^{(i)}) + d(x^{(i)}, C_1^{(i)}) \geqq 2R+2$$

(in the notation of § 1 of Part I). The construction begins by choosing $x^{(1)}, \cdots, x^{(n)}$.

Let B be any (linear or nonlinear) code of length n, minimal distance $d \geqq 6$, and containing at least n codewords. Let $x^{(1)}, \cdots, x^{(n)}$ be distinct codewords of B. We may assume (if necessary by complementing coordinates) that the ith coordinate of $x^{(i)}$ is 0, for $i = 1, \cdots, n$. Let

(61) $$S_i = \{y \in \mathbf{F}^n : d(y, x^{(i)}) \leqq [(d-2)/2] \text{ and the } i\text{th coordinate of } y \text{ is } 0\}$$

and define

(62) $$C = \mathbf{F}^n \backslash (S_1 \cup S_2 \cup \cdots \cup S_n).$$

THEOREM 19. *C has covering radius* 1 *and norm at least* $[d/2] + 1 \geqq 4$, *and is therefore abnormal.*

Proof. By the triangle inequality the sets S_i are disjoint. We first show C has covering radius 1. If $z \in \mathbf{F}^n$, $z \notin C$, then $z \in S_i$ for a unique i. Let c be obtained by changing the ith coordinate of z to 1. Then $d(z, c) = 1$, $c \notin S_i$ (by (61)), and $d(c, x^{(i)}) \leqq [d/2]$. Therefore $d(c, x^{(j)}) \geqq [d/2]$ for $j \neq i$, so $c \notin S_j$ and thus $c \in C$. Second, by construction $d(x^{(i)}, C_0^{(i)}) \geqq [d/2]$ and $d(x^{(i)}, C_1^{(i)}) \geqq 1$, and so the norm of C is at least $[d/2] + 1 \geqq 4 = 2R + 2$. Thus C is abnormal.

Example. Let B be a Hadamard code of length 11, and take the $x^{(i)}$ to be the eleven cyclic shifts of (01011100010). Then C is an abnormal code of length 11 containing 1432 codewords.

For a smaller example we use the vectors $x^{(1)}, \cdots, x^{(10)}$ of length 10 shown in Table 3, let $S_i = \{y \in \mathbf{F}^{10} : d(y, x^{(i)}) \leqq 2$, and ith coordinate of y is $0\}$, and define C by (62). Although the minimal distance between the $x^{(i)}$ is only 5 (so Theorem 19 does not

		$m_Q - 1$	
w_2'	1 0 0 · · · 0 1	1 1 1 1	w_2
x'	0 · · · · · · r	· · · ·	x
$b = b_0$	0 · · · · · · r	r r r r	a_0
b_1	1 · · · · · · r	r r r r	a_1

FIG. 4. *Vectors used in proof of Theorem* 17. *Their names in code B are shown in the left, and in A on the right.*

TABLE 3
Vectors $x^{(1)}, \cdots, x^{(10)}$ *used in construction of an abnormal nonlinear code.*

0	1	0	1	1	1	0	0	0	1
0	0	1	0	1	1	1	0	0	0
1	0	0	1	0	1	1	1	0	0
0	1	0	0	1	0	1	1	1	0
0	0	1	0	0	1	0	1	1	1
0	0	0	1	0	0	1	0	1	1
1	0	0	0	1	0	0	1	0	1
1	1	0	0	0	1	0	0	1	0
1	1	1	0	0	0	1	0	0	1
0	1	1	1	0	0	0	1	0	0

apply), C turns out to be an abnormal code of length 10, with 564 codewords and covering radius 1.

Finally, it is possible to omit many of those 564 codewords and still have an abnormal code. In this way we have constructed an abnormal code of length 10 with 217 codewords and covering radius 1 (we omit the details). This is the smallest abnormal code known at present. It would be interesting to find the smallest possible example.

Acknowledgments. We thank Peter Frankl for allowing us to include his construction in § 9, and J. S. Leon for allowing us to use his code automorphism program [5]. We also thank H. Janwa and H. F. Mattson, Jr. for their comments on a preliminary version of this paper. The first author thanks Bell Laboratories for its support and hospitality.

REFERENCES

[1] G. D. Cohen, M. G. Karpovsky, H. F. Mattson, Jr. and J. R. Schatz, *Covering radius—survey and recent results*, IEEE Trans. Inform. Theory, IT-31 (1985), pp. 328–343.

[2] G. D. Cohen, A. C. Lobstein and N. J. A. Sloane, *Further results on the covering radius of codes*, IEEE Trans. Inform. Theory, IT-32 (1986), pp. 680–694.

[3] R. L. Graham and N. J. A. Sloane, *On the covering radius of codes*, IEEE Trans. Inform. Theory, IT-31 (1985), pp. 385–401.

[4] H. J. Helgert and R. D. Stinaff, *Minimum-distance bounds for binary linear codes*, IEEE Trans. Inform. Theory, IT-19 (1973), pp. 344–356.

[5] J. S. Leon, *Computing automorphism groups of error correcting codes*, IEEE Trans. Inform. Theory, IT-28 (1982), pp. 496–511.

[6] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[7] H. F. Mattson, Jr., *An Improved Upper Bound on Covering Radius*, Lecture Notes Computer Science 228, 1986, pp. 90–106.

[8] D. Slepian, *Some further theory of group codes*, Bell System Tech. J., 39 (1960), pp. 1219–1252.

[9] N. J. A. Sloane, *Unsolved problems related to the covering radius of codes*, in Proc. Specific Problems in Communication and Computation (SPOC-85), Springer-Verlag, New York, 1987, to appear.

[10] ———, *A new approach to the covering radius of codes*, J. Combin. Theory Ser. A, 42 (1986), pp. 61–86.

[11] T. Verhoeff, *Updating a table of bounds on the minimum distance of binary linear codes*, Math. Dept., Eindhoven Univ. Tech., Report 85-WSK-01, Eindhoven, the Netherlands, January 1985; IEEE Trans. Inform. Theory, to appear.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

009

LATENT SUBSETS FOR DUAL INTERSECTING SYSTEMS*

P. C. FISHBURN†

Abstract. A collection F of subsets of $\{1, 2, \cdots, n\}$ is a *dual intersecting system* if no two sets in F have an empty intersection or an exhaustive union. Let f_j be the number of sets in F that contain point j, and let f_j^L be the number of sets not in F but included in some set in F that contain j.

We conjecture that if F is a dual intersecting system, then $f_j^L \geqq f_j$ for some j in $\{1, \cdots, n\}$. This is shown to be true if either $\min f_j \leqq n$ or $\min f_j < 8$.

Key words. intersecting systems, latent subsets

AMS(MOS) subject classifications. 05A05, 05C65

1. Introduction. This note discusses a conjecture for systems of subsets of a finite set that arose from research on related problems [2], including Chvátal's conjecture [1] and Kleitman's stronger conjecture [3].

A system F of subsets $A, B, \cdots$ of $\mathbf{n} = \{1, 2, \cdots, n\}$ is an *intersecting system* if $A \cap B \neq \emptyset$ for all $A, B \in F$ and a *dual intersecting system* if both F and its dual $F^* = \{\mathbf{n} \backslash A\colon A \in F\}$ are intersecting systems. The family of *latent subsets* for F is

$$F^L = \{A \subseteq \mathbf{n}\colon A \notin F, A \subset B \text{ for some } B \in F\}.$$

Let $F_j = \{A \in F\colon j \in A\}, f_j = |F_j|, F_j^L = \{A \in F^L\colon j \in A\}$ and $f_j^L = |F_j^L|$.

CONJECTURE. *For every $n \geqq 2$ and every dual intersecting system F of $\mathbf{n}$, $f_j^L \geqq f_j$ for some j in $\mathbf{n}$.*

Since a counterexample requires $f_j > f_j^L$ for all j, no generality is lost if we assume that $B \in F$ whenever $A \subset B \subset C$ and $A, C \in F$, and that F is a maximal intersecting system within $F \cup F^L$.

For a slightly different perspective on the conjecture, let $I_j = \{A \subseteq \mathbf{n}\colon j \in A\}$, the *principal filter* generated by $\{j\}$, and note that F^* is an intersecting system if and only if $A \cup B$ is a proper subset of $\mathbf{n}$ whenever $A, B \in F$. The conjecture says that if

(P1) $\quad A \cap B \neq \emptyset \quad$ for all $A, B \in F$;

(P2) $\quad A \cup B \neq \mathbf{n} \quad$ for all $A, B \in F$,

then $|F^L \cap I_j| \geqq |F \cap I_j|$ for some $j \leqq n$. Its resolution might aid in resolving the conjectures of Chvátal and Kleitman, which seem somewhat more intractable (see [2] for further discussion).

Partial confirmations of the conjecture appear in ensuing sections. Section 2 presents examples and suggests that the nearest one can get to a counterexample occurs when (P1) and (P2) operate independently in different parts of $\mathbf{n}$. Section 3 proves that the conjecture is true whenever $n \geqq \min f_j$. Section 4 discusses cases for small $\min f_j$ which show that the conjecture is true whenever $\min f_j < 8$.

The following known results will be used later.

LEMMA 1. *$|F| \leqq 2^{n-2}$ if F is a dual intersecting system of $\mathbf{n}$.*

LEMMA 2. *If F and G are systems of subsets of $\mathbf{n}$ with $A \cap B \neq \emptyset$ whenever $A \in F$ and $B \in G$, then*

$$|F^L||G^L| \geqq |F||G|.$$

* Received by the editors November 13, 1985; accepted for publication March 31, 1987.

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

Lemma 1 is proved in various places, the earliest of which appears to be [5]. Lemma 2 is the main result in Kleitman and Magnanti [4]. It clearly implies Corollary 1.

COROLLARY 1. $|F^L| \geqq |F|$ *for every intersecting system* F.

2. Examples. Figure 1 presents two maximal dual intersecting systems for $n = 6$ in terms of their characteristic vectors in $\{0, 1\}^6$. The system on the left has $|F| = 10$ and $|A| = 3$ for every A in F. Every other $B \subseteq \mathbf{6}$ is either disjoint from something in F (no matching 1's) or covers all of **6** in union with something in F (no matching 0's). Here $f_j^L = 6$ and $f_j = 5$ for all j.

The system on the right, with $|F| = 16$, is a maximum-cardinality dual intersecting system according to Lemma 1. It has $f_1 = f_2 = f_3 = 12$, $f_1^L = f_2^L = f_3^L = 4$, and $f_j = f_j^L$ for $j = 4, 5, 6$, so $f_j \geqq f_j^L$ for every j. It therefore fails as a counterexample only because of the equalities for $j \geqq 4$. Condition (P1) is wholly verified by the first three columns, while (P2) is established by the last three columns, so (P1) and (P2) operate independently on disjoint subsets of **6**.

Other maximum-cardinality dual intersecting systems follow the same pattern. If $\{A_1, A_2\}$ is a nontrivial partition of **n**, if G is a maximal intersecting system of A_1, and if H is a maximal union system of A_2 (no two sets in H have A_2 as their union), then

$$F = \{A \cup B: A \in G, B \in H\}$$

is a dual intersecting system with $|F| = |G||H| = 2^{|A_1|-1}2^{|A_2|-1} = 2^{n-2}$. In every such case, $f_j = f_j^L$ for each j in A_2, and $f_j > f_j^L$ for $j \in A_1$ if and only if j is contained in more than half of the sets in G.

I suspect that this type of example comes as close as possible to violating the conjecture. To be more precise and a bit more general, call a dual intersecting system F *separable* if there is a nontrivial partition $\{A_1, A_2\}$ of **n** such that $\{A_1 \cap B: B \in F\}$ is an

		j			
1	2	3	4	5	6
1	1	1	0	0	0
1	1	0	1	0	0
1	0	1	0	1	0
1	0	0	1	0	1
1	0	0	0	1	1
0	1	1	0	0	1
0	1	0	1	1	0
0	1	0	0	1	1
0	0	1	1	1	0
0	0	1	1	0	1

			j		
1	2	3	4	5	6
1	1	1	1	0	0
1	1	1	0	1	0
1	1	1	0	0	1
1	1	1	0	0	0
1	1	0	1	0	0
1	1	0	0	1	0
1	1	0	0	0	1
1	1	0	0	0	0
1	0	1	1	0	0
1	0	1	0	1	0
1	0	1	0	0	1
1	0	1	0	0	0
0	1	1	1	0	0
0	1	1	0	1	0
0	1	1	0	0	1
0	1	1	0	0	0

FIG. 1. *Maximal dual intersecting systems.*

intersecting system and no two sets in $\{A_2 \cap B\colon B \in F\}$ have A_2 as their union. Then it might be true that, if F is a nonseparable dual intersecting system, $f_j^L > f_j$ for some j.

An obvious special case of the following consequence of Corollary 1 says that every separated dual intersecting system satisfies the conclusion of the conjecture.

COROLLARY 2. *The conjecture is true for a dual intersecting system F of* **n** *if there is a j in* **n** *for which $F'_j = \{B\setminus\{j\}\colon B \in F_j\}$ is an intersecting system.*

Proof. Given the hypotheses for j, Corollary 1 says that $|F_j'^L| \geqq |F_j|$. Moreover, $f_j^L = |F_j'^L|$ and $f_j = |F'_j|$. □

3. Partial confirmation. The main purpose of this section is to prove Theorem 1.

THEOREM 1. *The conjecture is true for a dual intersecting system F of* **n** *if* $\min f_j \leqq n$.

We assume henceforth that F is a dual intersecting system of **n** and for definiteness let

$$c = f_1 = \min f_j.$$

The following consequence of Lemma 2 will be needed.

COROLLARY 3. *The conjecture is true for F if there are distinct j and k in* **n** *such that $\{j, k\}$ is not a subset of any set in F.*

Proof. Suppose that $\{j, k\} \cap A \neq \{j, k\}$ for all $A \in F$. Then $F'_j = \{B\setminus\{j\}\colon B \in F_j\}$ and $F'_k = \{B\setminus\{k\}\colon B \in F_k\}$ have $A \cap C \neq \varnothing$ whenever $A \in F'_j$ and $C \in F'_k$. Hence, by Lemma 2,

$$|F_j'^L||F_k'^L| \geqq |F'_j||F'_k|.$$

Suppose $|F_j'^L| \geqq |F'_j|$. Then $f_j^L \geqq f_j$ since $f_j^L = |F_j'^L|$ and $f_j = |F'_j|$. □

Since Theorem 1 is true even without $c \leqq n$ if the condition in Corollary 2 or Corollary 3 holds, we assume henceforth in this section that no F'_j is an intersecting system and that, for every 2-set $\{j, k\}$ in **n**, $\{j, k\} \subseteq A$ for some A in F. Figure 2 illustrates these requirements for $j = 1$ in a rearranged characteristic matrix of F.

We proceed under the assumption that $n \geqq c$.

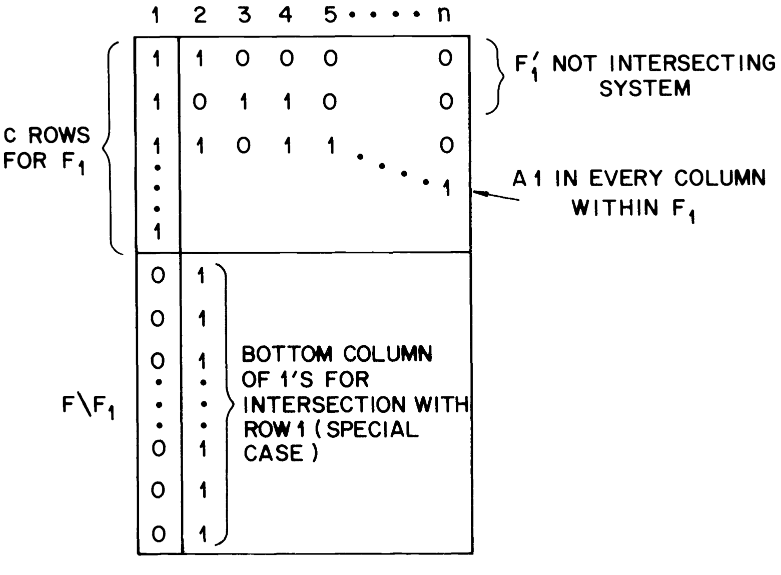

1 2 3 4 5 • • • • n
1 1 0 0 0 0
1 0 1 1 0 0
1 1 0 1 1 0
1
C ROWS FOR F1
F1′ NOT INTERSECTING SYSTEM
A 1 IN EVERY COLUMN WITHIN F1
F\F1
0 1
0 1
0 1
0 1
0 1
0 1
BOTTOM COLUMN OF 1'S FOR INTERSECTION WITH ROW 1 (SPECIAL CASE)

FIG. 2. *Characteristic matrix of F.*

If $|A| \geqq 2$ for every $A \in F'_1$, then $f_1^L \geqq c$ since $\{1\}, \{1, 2\}, \{1, 3\}, \cdots, \{1, n\}$ are in F_1^L.

Assume henceforth in this section that $|A| = 1$ for at least one $A \in F'_1$, and suppose initially that at least two $A \in F'_1$ are singletons, as shown for rows 1 and 2 in Fig. 3. Consider column 4 in Fig. 3 when $f_4 = m_1 + m_2$. Since each of the m_2 rows produces three distinct sets in F_4^L that begin 0101, 0011 and 0001, and since $00010 \cdots 0$ is another set in F_4^L, we get $f_4^L \geqq f_4$ unless $m_1 + m_2 > 3m_2 + 1$, or $m_1 \geqq 2(m_2 + 1)$. Since $f_4 \geqq c$ by the definition of c, $m_2 \geqq c - m_1$, and therefore

$$m_1 \geqq \tfrac{2}{3}(c+1).$$

Since $c - 2 \geqq m_1$, this implies $c \geqq 8$. A similar conclusion holds for each of columns 5–n. Therefore, if $f_j > f_j^L$ for all $j \geqq 4$, then there are at least $(n - 3)\lceil 2(c + 1)/3\rceil$ 1's in the double-lined submatrix of Fig. 3. It follows that there are at least

$$m = \lceil (n-3)\lceil 2(c+1)/3\rceil/(c-2)\rceil$$

1's in some row of that submatrix. This row alone produces (the worst case begins 1001) at least $2^m - (c - 2)$ sets in F_1^L, so $f_1^L \geqq f_1$ if $2^m - (c - 2) \geqq c$, or $2^m + 2 \geqq 2c$. Since this is easily seen to be true when $n \geqq c \geqq 8$, we conclude that the conjecture is true if $|A| = 1$ for more than one A in F'_1.

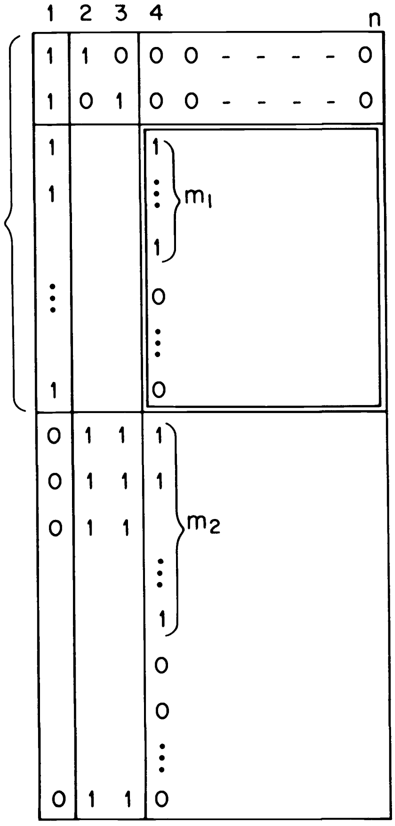


FIG. 3. $|A| = 1$ *for two* A *in* F'_1.

Assume henceforth in the proof of Theorem 1 that F'_1 contains exactly one singleton, say $\{2\}$. If $n > c$, then $f_1^L \geqq f_1$, so assume also that $n = c$ with $c = f_1 > f_1^L = c - 1$, where $F_1^L = \{\{1\}, \{1, 3\}, \cdots, \{1, c\}\}$.

Under these conditions, suppose there is no row in F_1 like row 2 in Fig. 4, i.e., no set of the form $\{1, j, k\}$ with $2 < j < k$ in F_1. Then each A in $F'_1 \backslash \{\{2\}\}$ must either have the form $\{2, j, \cdots\}$ or have $2 \notin A$ and $|A| \geqq 3$. If there are any sets of the latter type, then $f_1^L > c$, a contradiction, so all must contain 2. But then $2 \in B$ for every $B \in F$, so F'_j is an intersecting system for all $j \neq 2$, contrary to an earlier assumption. Hence our conditions assure an array like that shown in Fig. 4, except that some of the 10 or 01 entries in columns 3 and 4 in the lower part of the matrix can be 11. Since 11 there could only increase our computation of f_5^L in the next paragraph, we proceed with the array as shown.

The size of F_5^L must be at least $m_2 + m_3$ (0's in column 2) plus max $\{m_2, m_3\}$ (0's in column 3 or column 4) plus 2 ($\{5\}$, $\{1, 5\}$ from row 3). Therefore $f_5 > f_5^L$ requires $m_1 + m_2 + m_3 \geqq m_2 + m_3 + \max\{m_2, m_3\} + 3$. Since $f_5 \geqq c$, $m_2 + m_3 \geqq c - m_1$, so $\max\{m_2, m_3\} \geqq \lceil (c - m_1)/2 \rceil$. Then $m_1 \geqq \lceil (c - m_1)/2 \rceil + 3$, which is equivalent to

$$m_1 \geqq \lceil c/3 \rceil + 2.$$

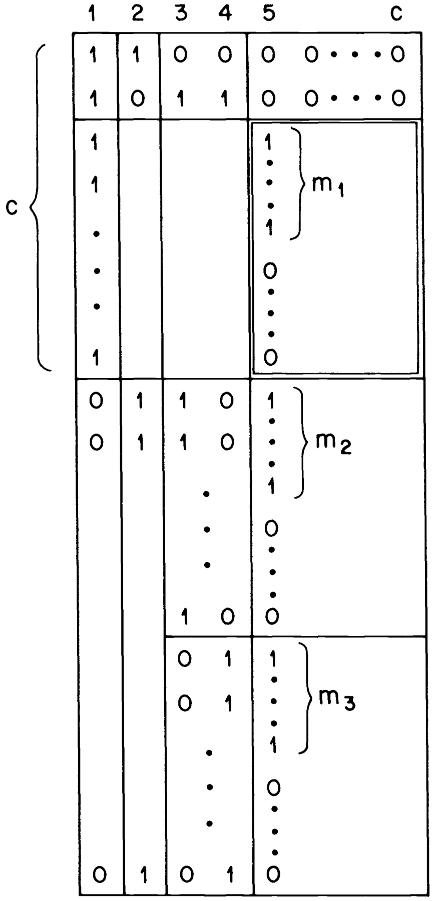


FIG. 4. *Special matrix for $n = c$.*

Since $c - 2 \geqq m_1$, this implies $c \geqq 6$. A similar result holds for every $j > 5$. If $c \in \{6, 7\}$ and $f_j > f_j^L$ for all $j \geqq 5$, the double-lined submatrix of Fig. 4 has no 0's and this yields $f_1^L \geqq f_1$ for a contradiction. On the other hand, if $c \geqq 8$ and $f_j > f_j^L$ for all $j \geqq 5$, then some row in the double-lined submatrix has at least

$$m = \lceil (c-4)(\lceil c/3 \rceil + 2)/(c-2) \rceil$$

1's. This row alone produces at least $2^m - (c - 2)$ sets in F_1^L (worst case begins 10001), and two more emerge from row 2, so $f_1^L \geqq 2^m + 4 - c$. However, $2^m + 4 - c \geqq c$ for all $c \geqq 8$, again contradicting $f_1 > f_1^L$.

Hence if $f_1 > f_1^L$ for Fig. 4, then $f_j^L \geqq f_j$ for some $j \geqq 5$. This completes the proof of Theorem 1.

4. Small min f_j. We continue to assume that F is a dual intersecting system of **n** with $F'_j = \{B \backslash \{j\} \colon B \in F\}$ and $c = f_1 = \min f_j$. The preceding section shows that if there is a counterexample to the conjecture then such an F must satisfy

(C1) No F'_j is an intersecting system;

(C2) For every 2-set $\{j, k\}$, $\{j, k\} \subseteq A$ for some $A \in F$;

(C3) $n < c$.

THEOREM 2. *If* $c \leqq 7$ *and* F *satisfies* (C1)–(C3), *then the conjecture is true for* F.

Proof. We prove Theorem 2 for $c = 7$. Proofs for $c < 7$ are similar and simpler.

The first part of the preceding proof with Fig. 3 shows that the conjecture is true for F when $n < c = 7$ if F'_1 contains more than one singleton, so assume henceforth that F'_1 has at most one singleton.

If $n = 5$, (C1) and (P2) imply that two rows of F are like the first two rows in Fig. 4. Then $f_5 > f_5^L$ forces $m_1 \geqq 5$, so to avoid $f_5^L \geqq f_5$ the fifth column of F must have 1's in rows 3–7. This then forces 0's in column 2 for these rows, so they all are $10{*}{*}1$. However, the $**$ positions cannot be filled with 0's and 1's in five different ways. Therefore $f_5^L \geqq f_5$.

Assume henceforth that $n = 6$. If two rows of F have the form shown at the top of Fig. 4 then, to avoid $f_j^L \geqq f_j$ for $j = 5$ or $j = 6$, we again get $m_1 = 5$, and the analysis in the preceding paragraph shows that this is impossible.

Assume henceforth that F'_1 does not contain a singleton and a doubleton that are disjoint. Suppose then that it has a singleton and tripleton that are disjoint (see Fig. 5). Condition (C2) then forces a row like row 3 in Fig. 5 with a 1 in column 6, a 0 in column 2 ((P2) with row 2) and exactly two 1's in columns 3–5. Since F can have at most two more rows in F_1 with 0's in column 2, it must have two more rows with 1's in column 2, hence 0's in columns 5 and 6 ((P2) with rows 2 and 3) and at least one 1 in columns 3 and 4 (see rows 4 and 5). These two rows eliminate 101011 and 100111 by (P2). But then we can add at most one more row in F_1 (say 111100 if the $*$'s are 0's), so F_1 cannot be completed.

Assume henceforth that F'_1 has no disjoint singleton and tripleton. Then (C1) implies that F'_1 has two disjoint doubletons. It then has no singleton, or else we would obtain the disjoint singleton/doubleton combination. Consequently, $f_1^L \geqq 6$ since $\{1\}$ and each $\{1, j\}$ are in F_1^L. It is easily seen that if F'_1 has two tripletons, say $\{2, 3, 4\}$ and $\{2, 3, 5\}$, then $f_1^L \geqq 7 = f_1$ unless all of 23, 24, 25, 34 and 35 are in F'_1. But then 6 is in no F_1 set, contrary to (C2).

We therefore arrive at the point for $(c, n) = (7, 6)$ where at least six sets in F'_1 are doubletons and the seventh is either a doubleton or a tripleton whose doubletons are

$$\begin{array}{c|c|ccccc|}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 1 & 1 & 1 & 0 \\
\hline
3 & 1 & 0 & 1 & 1 & 0 & 1 \\
4 & 1 & 1 & 1 & \ast & 0 & 0 \\
5 & 1 & 1 & \ast & 1 & 0 & 0 \\
6 & 1 & & & & & \\
7 & 1 & & & & & \\
\hline
 & 0 & & & & & \\
 & \vdots & & & & & \\
 & 0 & & & & & \\
\hline
\end{array}$$

FIG. 5. *A case for* $(c, n) = (7, 6)$.

$$\begin{array}{c|c|ccc|cc|}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & 1 & 1 & 1 & 1 & 0 & 0 \\
2 & 1 & 1 & 1 & 0 & 0 & 0 \\
3 & 1 & 1 & 0 & 1 & 0 & 0 \\
4 & 1 & 0 & 1 & 1 & 0 & 0 \\
\hline
5 & 1 & 0 & 0 & 1 & 1 & 0 \\
6 & 1 & & & & 1 & 0 \\
7 & 1 & & & & 0 & 1 \\
\hline
8 & 0 & & & & 0 & 1 \\
9 & 0 & & & & 0 & 1 \\
10 & 0 & & & & 0 & 1 \\
11 & 0 & & & & 0 & 1 \\
12 & 0 & & & & 0 & 1 \\
13 & 0 & & & & 0 & 1 \\
 & \vdots & & & & & \\
 & 0 & & & & & \\
\hline
\end{array}$$

FIG. 6. *Another case for* $(c, n) = (7, 6)$.

also in F'_1 to prevent $f_1^L \geqq 7$. Suppose F'_1 has a tripleton. Then F_1 must look like the upper part of Fig. 6 with some column, say column 6, having only one 1 in F_1. Because $f_6 \geqq c$, there must be at least seven rows in F that end in 01 since none can end in 11 ((P2) with row 1). Because of row 5 (F'_1 has all doubletons after row 1), each of rows 8–13 must have a 0 in column 2 or 3. To avoid duplication, one of these six rows would have to be 000001. But then (P1) is violated.

Finally, suppose all seven sets in F'_1 are doubletons, so its submatrix has exactly 14 1's. Then one of columns 2–6 has at most two 1's in F_1. When we examine the lower part of the F matrix for such a column in the five or more rows where it has a 1 ($f_j \geqq 7$), we find that it is impossible to complete these rows without violating (P1) or (P2) in reference to the upper part of the matrix. The straightforward details are left to the reader.

REFERENCES

[1] V. CHVÁTAL, *Intersecting families of edges in hypergraphs having the hereditary property*, Hypergraph Seminar, C. Berge and D. Ray-Chaudhuri, eds., Lecture Notes in Mathematics 411, Springer-Verlag, Berlin, 1974, pp. 61–66.

[2] P. C. FISHBURN, *Combinatorial optimization problems for systems of subsets*, SIAM Rev., 30 (1988), to appear.

[3] D. J. KLEITMAN, *Extremal hypergraph problems*, Surveys in Combinatorics, B. Bollobás, ed., Cambridge University Press, Cambridge, 1979, pp. 44–65.

[4] D. J. KLEITMAN AND T. L. MAGNANTI, *On the latent subsets of intersecting collections*, J. Combin. Theory Ser. A, 16 (1974), pp. 215–220.

[5] P. D. SEYMOUR, *On incomparable collections of sets*, Matematika, 20 (1973), pp. 208–209.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

010

EXTREMAL LENGTH AND WIDTH OF BLOCKING POLYHEDRA, KIRCHHOFF SPACES AND MULTIPORT NETWORKS*

SETH CHAIKEN†

Abstract. Various facts about the extremal length (EL) and extremal width (EW) of a one-port network on a Kirchhoff space due to Anderson, Duffin and Trapp and their relation to blocking pairs of polyhedra are unified and extended to the multiport case. The definitions of EL and EW are extended to all pairs of blocking polyhedra (G, H) on coordinates E given a symmetric positive definite matrix R. It follows that $\mathrm{EW}^{-1} = \min \{x^tRx \mid x \in G\}$, $\mathrm{EL}^{-1} = \min \{z^tR^{-1}z \mid z \in H\}$ and $\mathrm{EL}\cdot\mathrm{EW} = 1$. A Kirchhoff space on coordinates (E, P) where P is called the set of ports is a subspace that represents a matroid on $E \cup P$ in which P is independent and co-independent. Given any nonzero vector ω on port coordinates P, we extend Fulkerson's construction of a blocking pair from orthogonal subspaces with one distinguished coordinate to Kirchhoff spaces which model multiport networks. For ω and positive definite R a pair of minimization problems with reciprocal values are derived from Kirchhoff spaces. When R is diagonal these problems coincide with the EW^{-1} and EL^{-1} problems for the blocking pair from Kirchhoff spaces. In the case of a multiport resistor network, EW is the power dissipated when the voltage vector ω is applied to the ports.

Key words. polyhedral combinatorics, networks, multiports, length-width inequality, blocking polyhedra

AMS(MOS) subject classifications. 05, 15, 52, 90

1. Introduction. Moore and Shannon [MS] proved the length-width inequality for graphs which may be stated: Let G be a graph with two designated vertices s, t. Let $L(G)$ be the length of a shortest s-t path in G, and $W(G)$ be the number of edges in a smallest set that separates s and t. Then $L(G)W(G)$ is no greater than the number of edges in G.

Lehman [LEH] and Duffin [DUF62] extended the length-width inequality so that if l_i, w_i are arbitrary nonnegative weights assigned to each edge i of G, and

$$L(G) = \min_B \left(\sum_{i \in B} l_i \right),$$

$$W(G) = \min_A \left(\sum_{i \in A} w_i \right)$$

where B ranges over all s-t paths and A ranges over all s-t cuts, then

$$L(G)W(G) \leqq \sum_{i \in E(G)} l_i w_i.$$

Lehman also gave conditions on a system of sets equivalent to the property that the length-width inequality above holds with the system used in place of the s-t paths of a graph.

Duffin [DUF62] proved the length-width inequality for graphs using a technique borrowed from continuous analysis [DUF59]. He applied the notion of extremal width (EW) and extremal length (EL) of a quadrilateral with curved sides on which is defined a positive continuous "resistivity" function r to a 1-port electrical network in which each nonport edge i is given a positive resistance r_i. In both the curved quadrilateral and graph cases, EW and EL are defined by max-min optimizations (see § 2), $\mathrm{EW}\cdot\mathrm{EL} = 1$ and the

* Received by the editors June 10, 1985; accepted for publication April 10, 1987. An early version of this paper was supported by National Science Foundation grant MCS-8204796 and presented at the SIAM Second Conference on the Applications of Discrete Mathematics, Cambridge, Massachusetts, June 1983.

† Department of Computer Science, State University of New York, Albany, New York 12222.

solutions of the optimizations are electrical currents and potentials in the systems. Duffin also showed that EW·EL = 1 implies the length-width inequality. The physical significance of EW·EL = 1 in the curved quadrilateral case is that the equivalent resistance between one pair of opposite sides is the reciprocal of the equivalent resistance between the other pair of opposite sides [DUF62].

Let Z be the equivalent resistance of a 1-port resistive network. If current I is applied through the port the power dissipated is ZI^2. If, instead, potential difference E is applied across the port the power is $Z^{-1}E^2$. The fact that the power for unit current excitation is the reciprocal of the power for unit potential excitation is another manifestation of the EW·EL = 1 relationship. In §§ 4 and 5 a generalization of EW and EL is given for multiport networks. The EW·EL = 1 relationship for multiports is seen to generalize the above fact.

Anderson, Duffin and Trapp [ADT75], [AT77], [AT79] defined the linear algebraic concept of the Kirchhoff space or confluence to generalize the Kirchhoff law constraints on a multiport electrical network with no Kirchhoff law dependencies among the ports. In [AT77] they defined EL and EW for Kirchhoff spaces with one port coordinate. They proved EL·EW = 1 in this case and so demonstrated the length-width inequality for a situation involving orthogonal complimentary subspaces with one distinguished coordinate.

Fulkerson [FUL68], [FUL71], [SCHR] placed the length-width inequality into a general context in which he showed how it characterized that two polyhedra form a blocking pair (see Theorem 1). He also showed in [FUL68], [FUL70] how a pair of orthogonal complementary subspaces with one distinguished coordinate define a blocking pair of polyhedra on the other coordinates. Thus the length-width inequality for 1-ported Kirchhoff spaces was first established by Fulkerson apparently without the use of EL and EW.

In this paper we unify and extend the results that relate EL, EW, Kirchhoff spaces, and blocking pairs of polyhedra.

(1) We extend the definitions of EL, EW to all pairs of blocking polyhedra (G, H) on coordinates E. Given a positive definite symmetric matrix R on E, we show that Duffin's definitions [DUF62] are equivalent to $\mathrm{EW}^{-1} = \min \{x^tRx \mid x \in G\}$ and $\mathrm{EL}^{-1} = \min \{z^tR^{-1}z \mid z \in H\}$ and that EL·EW = 1 for all blocking polyhedra (G, H) (§ 2).

(2) Given a vector ω on the port coordinates, we show how a dual pair of Kirchhoff spaces defines a blocking pair of polyhedra (§ 3).

(3) Given a vector ω on the port coordinates of a dual pair of Kirchhoff spaces and a positive definite symmetric matrix R we give a pair of problems of the form $S = \min \{x^tRx\}$ and $T = \min \{z^tR^{-1}z\}$ for which $ST = 1$. When R is positive diagonal these problems are equivalent to result (1) applied to the blocking polyhedra from result (2) (§ 4).

(4) We observe that in multiport resistive electrical networks, EW is the power dissipated when the voltage vector ω is applied to the ports (§ 5).

In what follows, E and P denote disjoint finite sets of coordinate indices; for convenience let $E = \{1, 2, \cdots, n\}$. $\mathbb{R}^E$ is the set of tuples of real numbers indexed by E; $\mathbb{R}^E_+$ is the similar set of tuples of nonnegative reals. Inequalities and $| \quad |$ (absolute value) of vectors apply componentwise. 1 sometimes designates a vector of ones. $x \cdot z$ denotes the ordinary scalar product of vectors x and z.

2. Blocking polyhedra and their extremal length and width. We state the basic facts about blocking polyhedra (see [SCHR]).

Let G, H be the polyhedra

$$G=\{x \mid x \geqq 0, Ax \geqq 1\}, \qquad H=\{z \mid z \geqq 0, Bz \geqq 1\} \tag{2.1}$$

where A, B are nonnegative matrices with n columns, indexed by E. (G, H) are called a *blocking pair* when

$$H=\{z \mid z \geqq 0, x \cdot z \geqq 1 \text{ for all } x \in G\}.$$

THEOREM 1 (Fulkerson, Lehman). *The following conditions are equivalent:*

(2.2) *(G, H) is a blocking pair.*

(2.3) *H consists of all vectors z such that for some convex combination a of the rows $a_1, a_2, \cdots, a_l$ of A, $z \geqq a^t$. That is,*

$$H = \text{convex hull}\,(a_1, a_2, \cdots, a_l) + \mathbb{R}^n_+.$$

(2.4) *For all $w \geqq 0$,* $\min\{a_1 w, a_2 w, \cdots, a_l w\} = \max\{y \cdot 1 \mid y \geqq 0, yB \leqq w^t\}$.

(2.5) *$x \in G$, $z \in H$ implies that $x \cdot z \geqq 1$ and for all $l \geqq 0$, $w \geqq 0$*

$$\min_{x \in G}\{l \cdot x\} \min_{z \in H}\{w \cdot z\} \leqq l \cdot w.$$

(2.6) *(H, G) is a blocking pair.*

We note from (2.3) *that for all $l \geqq 0$ and $w \geqq 0$*

$$\min_{x \in G}\{l \cdot x\} = \min\{b_1 l, b_2 l, \cdots, b_k l\}, \quad \textit{and}$$

$$\min_{z \in H}\{w \cdot z\} = \min\{a_1 w, a_2 w, \cdots, a_l w\}.$$

Duffin [DUF62] defined EW and EL by (2.7) and (2.8) for positive diagonal R and what amount to the classic blocking pair (G, H) arising from s-t paths and cuts respectively in a network; also the alternative forms (2.9) and (2.10) were given and EW $\cdot$ EL $= 1$ was proved. In fact, the definitions make sense for all blocking pairs (G, H) and real positive definite symmetric matrices R:

$$\mathrm{EW} = \max_{x \in \mathbb{R}^n_+} \frac{(\min_{z \in H} x \cdot z)^2}{x^t R x}, \tag{2.7}$$

$$\mathrm{EL} = \max_{w \in \mathbb{R}^n_+} \frac{(\min_{x \in G} w^t R x)^2}{w^t R w}. \tag{2.8}$$

We will show later (Theorem 4) that for all blocking pairs and positive definite symmetric R, EW^{-1} and EL^{-1} are given also by Z and Y, respectively:

$$Z = \min_{x \in G}\{x^t R x\}, \tag{2.9}$$

$$Y = \min_{z \in H}\{z^t R^{-1} z\}. \tag{2.10}$$

First we analyze (2.9) and (2.10). The Cauchy–Schwarz inequality for positive definite symmetric R implies $(x^t R y)^2 \leqq (x^t R x)(y^t R y)$ for all x, y, with equality if and only if x and y are linearly dependent. The substitution $z = Ry$, $y = R^{-1}z$ gives

$(x\cdot z)^2 \leqq (x^tRx)(z^tR^{-1}z)$ with equality if and only if z and Rx are dependent. Hence $ZY \geqq 1$ for blocking (G, H). (This generalizes a proof in [DUF68].)

THEOREM 2. *Let (G, H) be a blocking pair of polyhedra as in Theorem* 1 *and let R be a symmetric positive definite matrix. Let Z and Y be given by* (2.9) *and* (2.10). *Then $ZY = 1$, solutions x_0, z_0 are unique and they satisfy $z_0 = YRx_0$.*

Proof. Using Theorem 1 and the positive definiteness of R, R^{-1} we pose the definitions of Z, Y as convex, quadratic programs. A has l rows and 1_l denotes a column vector of l ones.

$$\begin{aligned} Z &= \min x^tRx \\ &1_l - Ax \leqq 0 \\ &-x \leqq 0 \end{aligned} \tag{2.11}$$

$$\begin{aligned} Y &= \min z^tR^{-1}z \\ &z^t \geqq vA \\ &v \geqq 0 \\ &v1_l = 1. \end{aligned} \tag{2.12}$$

Let $L(x; \mu, \mu_1)$ denote the Lagrangian $x^tRx + \mu(1_l - Ax) + \mu_1(-x)$ for (2.11). Since, for example, Slater's condition that there exist x for which all inequalities are strict in (2.11) is satisfied (see [SW]), x_0 is a solution to (2.11) if and only if x_0 is feasible and there exist $\mu \geqq 0$, $\mu_1 \geqq 0$ that satisfy the following Kuhn–Tucker conditions:

$$\Delta_x L = 2x_0^tR - \mu A - \mu_1 = 0, \tag{2.13}$$

$$\mu(1_l - Ax_0) = 0, \tag{2.14}$$

$$\mu_1(-x_0) = 0. \tag{2.15}$$

We find $2x_0^tR \geqq \mu A$. By right multiplying $\Delta_x L = 0$ by x_0 and from (2.14) and (2.15) we derive $2x_0^tRx_0 = 2Z = \mu 1_l$. The vectors $v = (2Z)^{-1}\mu$ and $z_0^t = Z^{-1}x_0^tR$ are thus found to be feasible for program (2.12). Hence $Y \leqq z_0^tR^{-1}z_0 = Z^{-2}x_0^tRx_0 = Z^{-1}$. From $ZY \geqq 1$ we conclude that $ZY = 1$ and that $z_0^t = Z^{-1}x_0^tR = Yx_0^tR$ is a solution to (2.12). The uniqueness follows from the positive definiteness of R and R^{-1} and the convexity of the feasible sets in (2.11) and (2.12). □

If all the entries of R^{-1} are nonnegative the optimal solution z_0^t for (2.12) is given by $z_0^t = vA$ in the above proof because $vA \geqq 0$ is feasible and if $z^t \geqq vA \geqq 0$, then $z^tR^{-1}z \geqq z_0^tR^{-1}z_0$. In particular, this is true if R is an M matrix (see [VAR, p. 85]). From this, together with a similar argument for the dual, we conclude the following.

THEOREM 3. *If R is a nonnegative matrix the solution x_0 for* (2.9) *is given by $x_0^t = wB$, $w1_k = 1$, for some $w \geqq 0$. If R^{-1} is nonnegative the solution for* (2.10) *is given by $z_0 = vA$, $v1_l = 1$, for some $v \geqq 0$.*

We now prove the equivalence of the problems (2.9), (2.10) and (2.7), (2.8).

THEOREM 4. *Let (G, H) be a blocking pair and let R be a positive definite symmetric matrix. Assume* EW, EL *are defined by* (2.7), (2.8). *Then*

$$\mathrm{EW}^{-1} = \min_{x\in G} \{x^tRx\}, \tag{2.16}$$

$$\mathrm{EL}^{-1} = \min_{z\in H} \{z^tR^{-1}z\}. \tag{2.17}$$

Proof. It is sufficient to restrict x in (2.7) to $x \in \mathbb{R}_+^n$ for which $\min_{z \in H} x \cdot z \geqq 1$. This set of x is G by Theorem 1, (2.6); hence (2.16).

Let $K = \{Rw \mid w \in \mathbb{R}_+^n\}$. The substitution $z = Rw$ puts (2.8) into the form

$$\text{(2.18)} \qquad \mathrm{EL} = \max_{z \in K} \frac{(\min_{x \in G} z \cdot x)^2}{z^t R^{-1} z}.$$

Since R is positive definite, every nonzero vector in K has at least one positive component. We show that the maximization in (2.18) can be restricted to nonnegative z by proving that for each $z \in K$ with a negative component there exists $x \in G$ with $z \cdot x = 0$. Let $S = \{i \mid z_i < 0, 1 \leqq i \leqq n\}$ and $T = \{i \mid z_i > 0, 1 \leqq i \leqq n\}$. A strictly positive $x_0 \in \mathbb{R}_n^+$ with $x_0 \cdot z = 0$ is given by

$$(x_0)_j = \begin{cases} \sum\limits_{i \in T} z_i & \text{if } j \in S, \\ -\sum\limits_{i \in S} z_i & \text{if } j \in T, \\ 1 & \text{otherwise.} \end{cases}$$

Theorem 1 implies that (2.3) applies to G. Therefore $x = Nx_0$ is in G for sufficiently large $N > 0$.

Let $K_+ = K \cap \mathbb{R}_+^n$ so that

$$\mathrm{EL} = \max_{z \in K_+} \frac{(\min_{x \in G} z \cdot x)^2}{z^t R^{-1} z} \leqq \max_{z \in \mathbb{R}_+^n} \frac{(\min_{x \in G} z \cdot x)^2}{z^t R^{-1} z} = Y^{-1}.$$

The last equality is similar to (2.16). Now by Theorem 2 a solution $z_0 = YRx_0$ for the right-hand maximization is in K_+. We therefore have $\mathrm{EL} = Y^{-1}$. □

For positive diagonal R the equivalence (2.17) is almost immediate [DUF62]. In the same paper, $\mathrm{EL} \cdot \mathrm{EW} = 1$ with (2.7), (2.8) was shown to imply the length-width inequality (2.5). Note we cannot use this result to prove the length-width inequality for general blocking pairs because we used Theorem 1 to prove $\mathrm{EL} \cdot \mathrm{EW} = 1$.

3. Blocking polyhedra from Kirchhoff spaces. In this section we generalize results of Fulkerson [FUL68], [FUL70]: Let D, $D^\perp$ be complementary orthogonal subspaces of $\mathbb{R}^{n+1}$ under the ordinary scalar product. Assume that if $(u_E, 1) \in D$ then $u_E \neq 0$ and $(u_E, 1) \in D$ for some u_E. Let

$$\text{(3.1)} \qquad G = \{x \in \mathbb{R}_+^n \mid x_i \geqq |u_i| \, i \in E \text{ for some } (u_E, 1) \in D\},$$

$$\text{(3.2)} \qquad H = \{z \in \mathbb{R}_+^n \mid z_i \geqq |e_i| \, i \in E \text{ for some } (e_E, 1) \in D^\perp\}.$$

For all $x \in G$ $z \in H$, $x \cdot z \geqq \sum |u_i| |e_i| \geqq |\sum u_i e_i| = |1| = 1$. Fulkerson showed, in fact, that (G, H) is a blocking pair. Our results show that Fulkerson's construction also produces blocking pairs that arise from subspaces that model Kirchhoff's laws type of constraints on multiport networks.

Let $\mathbf{E}$ and $\mathbf{P}$ be finite-dimensional complex vector spaces with hermitian product $\langle \; , \; \rangle$. Let $\mathbf{V} = \mathbf{E} \oplus \mathbf{P}$. Anderson and Trapp say a subspace C of $\mathbf{V}$ is a confluence

[AT79] or Kirchhoff space [AT77] provided that

(3.3) for all $u_P \in \mathbf{P}$ there is a $u_E \in \mathbf{E}$ so that $(u_E, u_P) \in C$, and

(3.4) if $(0, u_P) \in C$, then $u_P = 0$.

They define the dual Kirchhoff space $C^\perp$ by

$$C^\perp = \{(e_E, e_P) | \langle e_E, u_E \rangle = \langle e_P, u_P \rangle \text{ for all } (u_E, u_P) \in C\}.$$

They prove $C^\perp$ is a Kirchhoff space whenever C is. C and $C^\perp$ represent, for example, the spaces of currents and voltages allowed by Kirchhoff's laws in an electrical network with ports P in which the interconnection allows any combination of currents u_P to be applied through the ports and in which it also allows any combination of voltages e_P to be applied to the ports. In the following C, $C^\perp$ will represent a dual pair of Kirchhoff spaces.

For our purposes, we restrict $\mathbf{V}$ to be the real vector space with coordinates $E \cup P$; we write $\mathbf{V} = R^{E \cup P}$, $\mathbf{E} = R^E$, $\mathbf{P} = R^P$. Let M, M^* be the dual pair of matroids [WEL], [WHI] on $(E \cup P)$ whose cycle space representations are C, $C^\perp$. The conditions (3.3) and (3.4) become the following:

(3.5) E contains a base for M, i.e., P is co-independent in M, and

(3.6) P is independent in M.

In this form the conditions that C be a Kirchhoff space are clearly self dual.

THEOREM 5. *Let $\omega \in \mathbb{R}^P$, $\omega \neq 0$ and C, $C^\perp$ be a dual pair of Kirchhoff spaces. The polyhedra*

(3.7) $G = \{x \in \mathbb{R}^E_+ | x_i \geqq |u_i| \, i \in E$ *for some* $(u_E, u_P) \in C$ *with* $u_P \cdot \omega = 1\}$,

(3.8) $H = \{z \in \mathbb{R}^E_+ | z_i \geqq |e_i| \, i \in E$ *for some* $(e_E, \omega) \in C^\perp\}$

are a pair of blocking polyhedra.

Proof. Given C, $C^\perp$ consider the subspaces of $\mathbb{R}^{n+1}$:

(3.9) $$D = \{(u_E, u_P \cdot \omega) | (u_E, u_P) \in C\},$$

(3.10) $$D^\perp = \{(e_E, -k) | (e_E, k\omega) \in C^\perp, k \text{ is real}\}.$$

Let D_1 temporarily denote the right-hand side of (3.10). We will show that $D_1 = D^\perp$ where $D^\perp$ denotes the orthogonal complement of D in $\mathbb{R}^{n+1}$ under the ordinary scalar product $(u_E, u_{n+1}) \cdot (e_E, e_{n+1}) = u_E \cdot e_E + u_{n+1} e_{n+1}$. It is easy to verify that D and D_1 are subspaces. Since $(u_E, u_P \cdot \omega) \cdot (e_E, -k) = u_E \cdot e_E - k u_P \cdot \omega = u_E \cdot e_E - u_P \cdot (k\omega)$ and C, $C^\perp$ are dual Kirchhoff spaces, the vectors in D are all orthogonal to the vectors in D_1, so $D_1 \subseteq D^\perp$. If $(e_E, -k) \in D^\perp$, then the same calculation shows that for every $(u_E, u_P) \in C$, $u_E \cdot e_E - u_P \cdot (k\omega) = 0$ so $(e_E, k\omega) \in C^\perp$ and hence $(e_E, -k) \in D_1$. We conclude that $D_1 = D^\perp$. It is now clear that G and H in (3.7), (3.8) are of the form (3.1), (3.2) and so they are a blocking pair. □

To prove G, H in (3.1), (3.2) were a blocking pair Fulkerson explicitly constructed the blocking matrices A, B that define G, H by (2.1) and then he verified condition (2.4), the "max flow-min cut" identity, in Theorem 1. We will provide here the generalizations of Fulkerson's notions for describing the blocking matrices that define G, H for Kirchhoff spaces given ω given by Theorem 5.

Let u be a vector. The *support* $S(u)$ is the set of coordinates i such that $u_i \neq 0$. If C is a vector space then $u \in C$ is said to be *elementary* in C if $u \neq 0$ and for no nonzero

$u' \in C$ does $S(u)$ properly contain $S(u')$. If u and u' are elementary in C and $S(u) = S(u')$, then u is a nonzero multiple of u'. It follows that there is only a finite set of vectors $\{(u^j_E, 1) | j = 1, \cdots, k\}$ that are elementary in D of the form $(u_E, 1)$ and there is only a finite set $\{(e^j_E, -1) | j = 1, \cdots, l\}$ of elementary vectors in $D^\perp$ of the form $(e_E, -1)$. Fulkerson defined A to be the $l \times n$ matrix whose jth row is $(|e^j_1|, |e^j_2|, \cdots, |e^j_n|)$. B is the $k \times n$ matrix whose jth row is $(|u^j_1|, |u^j_2|, \cdots, |u^j_n|)$.

DEFINITION. $(e_E, \omega) \in C^\perp$ is *ω-elementary* in $C^\perp$ if it is nonzero and $S(e_E)$ is minimal along $S(e'_E)$ for which $(e'_E, \omega) \in C^\perp$. $(u_E, u_P) \in C$ is *P-elementary* in C if it is nonzero and $S(u_E)$ is minimal among $S(u'_E)$ for which there exists nonzero $(u'_E, u'_P) \in C$.

Properties (3.3) and (3.4) of Kirchhoff space $C^\perp$ imply that for every $\omega \neq 0$ there exist ω-elementary vectors in $C^\perp$ and that $(0, \omega)$ is never one of them because $(0, \omega) \notin C^\perp$. The same properties applied to C imply there exist P-elementary vectors in C and $u_E \neq 0$ for every one of them.

PROPOSITION 6. *Suppose $C, C^\perp$ are a dual pair of Kirchhoff spaces; $D, D^\perp$ are given by* (3.9), (3.10) *and* $\{(u^j_E, 1)\}$ *and* $\{(e^j_E, -1)\}$ *are the elementary vectors in $D, D^\perp$ of the form* $(u_E, 1)$, $(e_E, -1)$, *respectively. Then* $\{(u^j_E, 1)\}$ *is the set of vectors* $(u_E, 1)$ *for which* (u_E, u_P) *is P-elementary in C and* $u_p \cdot \omega = 1$ *for some* u_p. $\{(e^j_E, -1)\}$ *is the set of vectors* $(e_E, -1)$ *for which* (e_E, ω) *is ω-elementary in* $C^\perp$.

Proof. We show that if (u_E, u_P) is P-elementary in C and $u_P \cdot \omega = 1$, then $(u_E, 1)$ is elementary in D and that if (e_E, ω) is ω-elementary in $C^\perp$, then $(e_E, -1)$ is elementary in $D^\perp$. The remainder of the proof is straightforward.

Suppose (u_E, u_P) is P-elementary in C and $u_P \cdot \omega = 1$. $(u_E, u_P \cdot \omega) = (u_E, 1) \in D$ and we must show that $(u_E, 1)$ is elementary in D. Suppose not; so there is a $(u'_E, k) = (u'_E, u'_P \cdot \omega) \in D$ with $S(u'_E, k) \subsetneqq S(u_E, 1)$. If $k \neq 0$, then $(u'_E, u'_P) \in C$ and $S(u'_E) \subsetneqq S(u_E)$. If $k = 0$, it is possible that $S(u'_E) = S(u_E)$. In that case, choose $i \in E$ so that $u_i \neq 0$ and let $(u''_E, u''_P) = (u_E, u_P) - (u_i/u'_i)(u'_E, u'_P)$. Then $(u''_E, u''_P) \in C$; it is nonzero because $u_P \cdot \omega = 1$ and $u'_P \cdot \omega = 0$, and $S(u''_E) \subseteq S(u_E) - \{i\} \subsetneqq S(u_E)$. In either case we contradict the assumption that (u_E, u_P) is P-elementary.

Now suppose (e_E, ω) is ω-elementary in $C^\perp$ and suppose $(e_E, -1) \in D^\perp$ is not elementary in $D^\perp$. Then $S(e'_E, k) \subsetneqq S(e_E, -1)$ for some $(e'_E, k) \in D^\perp$. If $k \neq 0$, then $(-k^{-1})(e'_E, -k\omega) = (e''_E, \omega) \in C^\perp$ with $S(e''_E) = S(e') \subsetneqq S(e_E)$. If $k = 0$, then $(e'_E, 0) \in C^\perp$. Let $e'_i \neq 0$ and $(e''_E, \omega) = (e_E, \omega) - (e_i/e'_i)(e'_E, 0)$. Again it is a contradiction that (e_E, ω) is ω-elementary. □

We conclude that there are finitely many P-elementary (u_E, u_P) in C with $u_P \cdot \omega = 1$ and that there are finitely many ω-elementary vectors in $C^\perp$. Property (3.4) of C implies that for each u_E above there is a unique u_P for which $(u_E, u_P) \in C$. From Fulkerson's descriptions of A and B in [FUL70] we conclude the following.

COROLLARY 7. *The blocking matrices A, B that define the polyhedra G, H* ((3.7), (3.8) *in Theorem* 5) *by* (2.1) *are as follows. A is the $l \times n$ matrix whose jth row is* $(|e^j_1|, |e^j_i|, \cdots, |e^j_n|)$ *where* $\{(e^j_E, \omega) | 1 \leqq j \leqq l\}$ *is the set of ω-elementary vectors in* $C^\perp$. *B is the $k \times n$ matrix whose jth row is* $(|u^j_1|, |u^j_2|, \cdots, |u^j_n|)$ *where*

$$\{(u^j_E, u^j_P) | 1 \leqq j \leqq k\}$$

is the set of P-elementary vectors in C for which $u^j_P \cdot \omega = 1$.

4. Extremal length and width of a dual pair of Kirchhoff spaces. Theorem 5 reveals a relationship between the minimizations (2.9) and (2.10) in Theorem 2 and a variational formulation of resistive electrical network problems. Let $C, C^\perp$ be a dual pair of Kirchhoff spaces, $\omega \in R^P$, $\omega \neq 0$. Let R be a positive definite symmetric matrix. Consider the minimization problems:

$$S = \min \{u_E^t R u_E | (u_E, u_P) \in C, u_P \cdot \omega = 1\}, \tag{4.1}$$

$$T = \min \{e_E^t R^{-1} e_E | (e_E, \omega) \in C^\perp\}. \tag{4.2}$$

THEOREM 8. *u_E minimizes* (4.1) *above if and only if u_E is a solution to*

$$(u_E, u_P) \in C, \quad u_P \cdot \omega = 1, \quad (Ru_E, k\omega) \in C^\perp \quad (k \textit{ is real}) \tag{4.3}$$

in which case $S = k$.

Proof. By the semi-definiteness of R, u_E solves the minimization if and only if $(u_E, u_P) \in C$, $u_P \cdot \omega = 1$ and the variation $\delta(u_E^t R u_E) = 2(u_E^t R)\delta u_E = 2(Ru_E) \cdot \delta u_E$ vanishes whenever δu_E is such that for some δu_P, $(\delta u_E, \delta u_P) \in C$ and $\delta u_P \cdot \omega = 0$. We show this $\delta(u_E^t R u_E) = 0$ is equivalent to $(Ru_E, k\omega) \in C^\perp$.

Let $C_1 = \{(u_E, u_P) \in C | u_P \cdot \omega = 0\}$. By condition (3.3) that C is a Kirchhoff space, $\dim C_1 = \dim C - 1$. Hence

$$C_1^\perp = \{(e_E, e_P) | (e_E, e_P) \cdot (u_E, u_P) = 0 \text{ for all } (u_E, u_P) \in C_1\}$$

has dimension $\dim C^\perp + 1$. Let

$$C_2 = \{(e_E, e_P + k\omega) | (e_E, e_P) \in C^\perp, k \text{ is real}\}.$$

C_2 has dimension $\dim C^\perp + 1$ also by (3.4) applied to Kirchhoff space $C^\perp$. Hence to show $C_2 = C_1^\perp$, it suffices to show that $C_2 \subseteq C_1^\perp$. Let $(e_E, e_P + k\omega) \in C_2$ and $(u_E, u_P) \in C_1$. It follows that

$$(e_E, e_P + k\omega) \cdot (u_E, u_P) = (e_E, e_P) \cdot (u_E, u_P) - k\omega \cdot u_P = 0$$

because $(e_E, e_P) \in C^\perp$, $(u_E, u_P) \in C$ and $\omega \cdot u_P = 0$.

Let $D = \{\delta u_E | (\delta u_E, \delta u_P) \in C_1\}$ and $D^\perp = \{e_E \in \mathbb{R}^E | e_E \cdot \delta u_E = 0 \text{ for all } \delta u_E \in D\}$; hence the condition that $\delta(u_E^t R u_E)$ vanish is $Ru_E \in D^\perp$. Let $e_E = Ru_E$.

$$\begin{aligned} e_E \in D^\perp &\Leftrightarrow \text{for all } (\delta u_E, \delta u_P) \in C_1, (\delta u_E, \delta u_P) \cdot (e_E, 0) = 0 \\ &\Leftrightarrow (e_E, 0) \in C_1^\perp = C_2 \\ &\Leftrightarrow (e_E, k\omega) \in C^\perp \text{ for some real } k. \end{aligned}$$

From (4.3), $(u_E, u_P) \cdot (Ru_E, k\omega) = u_E^t R u_E - k u_P \cdot \omega = S - k = 0$. □

The result dual to Theorem 8 dates to Rayleigh and Maxwell, see [DUF59]. Its Kirchhoff space formulation below was given by Anderson and Trapp.

THEOREM 9 [AT79]. *e_E minimizes* (4.2) *above if and only if e_E is a solution to*

$$(R^{-1} e_E, u_P) \in C, \quad (e_E, \omega) \in C^\perp \tag{4.4}$$

in which case $T = u_P \cdot \omega$.

These authors used the existence and uniqueness of solutions to (4.4) for all ω and positive real R^{-1} to define a linear operator μ so $\mu(\omega) = u_P$. Thus $T = \mu(\omega) \cdot \omega$. A fixed Kirchhoff space defines an operator Φ that associates by (4.4) the operator $\mu = \Phi(R^{-1})$ to each positive real R^{-1}. Φ was used by Anderson, Duffin and Trapp to study matrix operations based on interconnection of networks; see for example [ADT75], [AT79]. See also [BD53] for a generalization of (4.4) in which the solution is characterized as the stationary point of a quadratic function.

COROLLARY 10. *If S and T are given by* (4.1), (4.2), *then $ST = 1$.*

Proof. e_E, u_P is a solution to (4.4) if and only if $u'_E = R^{-1} e_E / u_P \cdot \omega$, $u'_P = u_P / u_P \cdot \omega$ and $k = 1/u_P \cdot \omega$ comprise a solution to (4.3). Hence $S = 1/T$. □

When R is a positive diagonal matrix $R = \operatorname{diag}(r_1, r_2, \cdots, r_n)$ and G, H are as in Theorem 5 the problem (2.9) is equivalent to (4.1) and (2.10) is equivalent to (4.2). Specifically, $u_E^t R u_E = \sum_{i=1}^n r_i u_i^2 = \sum_{i=1}^n r_i |u_i|^2$ and $x^t R x > x'^t R x'$ whenever $x_i \geqq x'_i \geqq 0$ for $1 \leqq i \leqq n$ with $x_i > x'_i$ for at least one i. Similarly for R^{-1}. Consequently,

$$\mathrm{EW}^{-1} = \min \{x^t R x \mid x \in G\} = \min \{u_E^t R u_E \mid (u_E, u_P) \in C, u_P \cdot \omega = 1\},$$

$$\mathrm{EL}^{-1} = \min \{z^t R^{-1} z \mid z \in H\} = \min \{e_E^t R^{-1} e_E \mid (e_E, \omega) \in C^\perp\}.$$

When positive definite R is not diagonal a transformation may be performed on the E coordinates of C to diagonalize R in (4.1) and so we obtain an EW^{-1} problem and its dual over blocking pairs.

5. Examples and application to electrical networks.

Example of Theorem 2. Let $n = 2$, G be defined by $x_1 \geqq a \geqq 0$, $x_2 \geqq b \geqq 0$ (Fig. 1) and

$$R = \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & 1 \end{bmatrix}, \qquad R^{-1} = \frac{1}{1-\varepsilon^2} \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{bmatrix} \tag{5.1}$$

where $0 \leqq \varepsilon < 1$. Without loss of generality assume $b \leqq a$. When $0 \leqq b/a < \varepsilon < 1$ the minimum (2.9) is attained by $x^t = (a, \varepsilon a)$ with $Z = a^2(1 - \varepsilon^2)$. When $0 \leqq \varepsilon \leqq b/a$ the minimum is attained by the vertex $x^t = (a, b)$ with $Z = a^2 - 2\varepsilon ab + b^2$. The blocker H of G is defined by $az_1 + bz_2 \geqq 1$, $z_1 \geqq 0$, $z_2 \geqq 0$ (Fig. 1). When $0 \leqq b/a < \varepsilon < 1$, (2.10) is minimized with $Y = a^{-1}(1 - \varepsilon^2)^{-1}$ by $z = Z^{-1}Rx = (a^{-1}, 0)^t$ which illustrates Theorem 2. When $\varepsilon \leqq b/a$ the minimizing z^t is $Z^{-1}(a - \varepsilon b, b - \varepsilon a)$.

We now illustrate that despite Theorem 5 and Corollary 10, when R is not diagonal, there is no general relationship between S (4.1) and Z (2.9) where G is defined from C and ω by Theorem 5 (3.7). Note (4.1) is a minimization over an affine subspace A. Let $n = 2$. First, let $A = \{(1, 0)^t\}$ and R be (5.1). $S = 1$ trivially. However, G is defined by $x_1 \geqq 1$, $x_2 \geqq 0$; see the previous example with $(a, b) = (1, 0)$ so that $Z = 1 - \varepsilon^2 < S$; the minimizing x is $(1, \varepsilon)^t$. Next, let $A = \{(1, k)^t \mid k \text{ is real}\}$ and R be (5.1) with $-1 < \varepsilon < 0$. Equation (4.1) is minimized by $u_E = (1, \varepsilon)$ and $S = 1 - \varepsilon^2$ but the minimum for G now occurs at $x^t = (1, 0)$ so $Z = 1 < 1 - \varepsilon^2 = S$. Of course an example with $Z < S$ is immediate

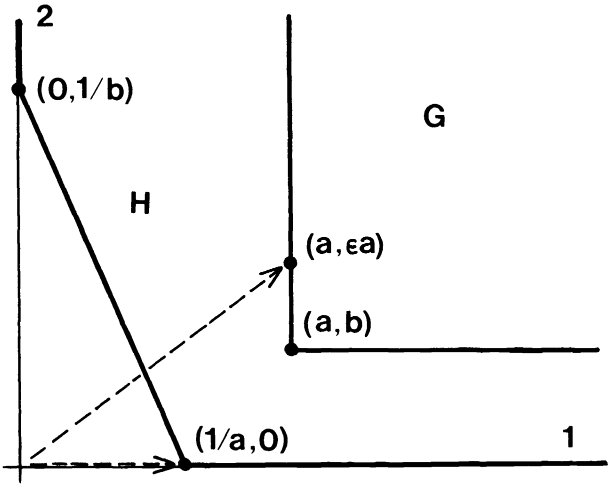


FIG. 1. $G = \{(a, b)^t\} + R_+^2$. H = *convex hull* $\{(a^{-1}, 0)^t, (0, b^{-1})^t\} + R_+^2$.

from the example with $Y > T$ which exists as a consequence of Theorem 2 and Corollary 10.

We give two dual electrical network interpretations to EW and EL by way of Theorems 8 and 9 in the case of positive diagonal R. For further information on electrical interpretations and their origins in distributed networks see [DUF59] and [DUF62].

(1) Let C' be the cycle space of the graph N with edges $E \cup P$ and $C^{\perp}$ be the cocycle space. $C = \{(u_E, u_P) | (u_E, -u_P) \in C'\}$. Suppose that each $i \in E$ is a resistor in an electrical network with resistance r_i, and each edge in P is a port.

For $i \in E$, u_i represents the current in edge i. For $i \in P$, u_i represents the current supplied to port edge i by an external connection. For $i \in E \cup P$, e_i represents the potential difference or voltage across edge i.

$(u_E, u_P) \in C$ is the condition that the edge currents satisfy Kirchhoff's current law. $(e_E, e_P) \in C^{\perp}$ is the condition that the potential difference e_E across resistors and the potential differences e_P supplied to the ports satisfy Kirchhoff's voltage law. The significance of $(u_E, u_P) \cdot (e_E, e_P) = u_E \cdot e_E - u_P \cdot e_P = 0$ is that the power dissipated in the network equals the power supplied. $u_E = e_E R^{-1}$ is Ohm's law. Therefore (4.4) represents the situation where potential (voltage) sources ω are connected to the ports. u_P represents the currents that flow into the ports in response and e_E represents the potential differences that appear across the resistors. $\mathrm{EL}^{-1} = \mathrm{EW} = T = u_P \cdot \omega$ is the power dissipated in this situation.

The condition $u_P \cdot \omega = 1$ in (4.3) is equivalent to $u_P \cdot (k\omega) = k$. Since $k = S = T^{-1}$, the solution to (4.3) is interpreted as the currents and potentials in the same network after the potential sources have been adjusted by a proportionality constant k until the power dissipated becomes T^{-1}. As noted already in the Introduction, for a single port network and $\omega = 1$ this adjustment is achieved when the port current becomes 1.

(2) Let C be the cocycle space of the graph N and C' be the cycle space. $C^{\perp} = \{(e_E, -e_P) | (e_E, e_P) \in C'\}$. Suppose each $i \in E$ is a resistor with resistance r_i^{-1}, and each edge in P is a port. We interchange the meanings of e and u from 1 so now e represents potential differences and u represents currents. Therefore (4.4) now represents the situation where current sources ω are connected to the ports. Again $\mathrm{EL}^{-1} = \mathrm{EW} = T = u_P \cdot \omega$ is the power dissipated.

Acknowledgments. The author wishes to thank the referees for comments that led to stronger results and better proofs. We also thank Sally Goodall for word processing of the manuscript and revisions.

REFERENCES

[ADT75] W. N. Anderson, R. J. Duffin and G. E. Trapp, *Matrix operations induced by network connections*, SIAM J. Control, 13 (1975), pp. 446–461.

[AT77] W. N. Anderson and G. E. Trapp, *Algebraic properties of networks on matroids*, Proc. 20th Midwest Symposium Circuits and Systems, Lubbock, TX, 1977, pp. 384–390.

[AT79] ———, *Matrix operations induced by electrical network connections—a survey*, in Constructive Approaches to Mathematical Models, C. V. Coffman and G. J. Fix, eds., Academic Press, New York, 1979, pp. 53–73.

[BD53] R. Bott and R. J. Duffin, *On the algebra of networks*, Trans. Amer. Math. Soc., 74 (1953), pp. 99–109.

[DUF59] R. J. Duffin, *Distributed and lumped networks*, J. Math. Mechanics, 8 (1959), pp. 793–826.

[DUF62] ———, *The extremal length of a network*, J. Math. Anal. Appl., 5 (1962), pp. 200–215.

[DUF68] ———, *Optimum heat transfer and network programming*, J. Math. Mechanics, 17 (1968), pp. 759–768.

[FUL68] D. R. FULKERSON, *Networks, frames and blocking systems*, in Mathematics of the Decision Sciences, Lectures in Applied Mathematics 11, G. B. Dantzig and A. F. Veinott, eds., American Mathematical Society, Providence, RI, 1968, pp. 303–334.

[FUL70] ———, *Blocking polyhedra*, in Graph Theory and its Applications, B. Harris, ed., Academic Press, New York, 1970, pp. 93–112.

[FUL71] ———, *Blocking and anti-blocking pairs of polyhedra*, Math. Programming, 1 (1971), pp. 168–194.

[LEH] A. LEHMAN, *On the length width inequality*, Math. Programming, 17 (1979), pp. 403–416 (orig. written 1963).

[MS] E. F. MOORE AND C. E. SHANNON, *Reliable circuits using less reliable relays*, J. Franklin Inst., 262 (1956), pp. 191–208, 281–298.

[SCHR] A. SCHRIJVER, *Fractional packing and covering*, in Packing and Covering in Combinatorics, Math. Centre Tract 106, Mathematisch Centrum, Amsterdam, 1979, pp. 201–274; *Some background from linear algebra*, in Packing and Covering in Combinatorics, Math. Centre Tract 106, Mathematisch Centrum, Amsterdam, 1979, pp. 5–15.

[SW] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions* I, Springer-Verlag, Berlin, Heidelberg, 1970.

[VAR] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[WEL] D. J. A. WELSH, *Matroid Theory*, Academic Press, London, 1976.

[WHI] N. WHITE, *Theory of Matroids*, Cambridge University Press, Cambridge, 1986.

A NEW HEURISTIC FOR MINIMUM WEIGHT TRIANGULATION*

ANDRZEJ LINGAS†

Abstract. A new heuristic for minimum weight triangulation of planar point sets is proposed. First, a polygon whose vertices are all points from the input set is constructed. Next, a minimum weight triangulation of the polygon is found by dynamic programming. The union of the polygon triangulation with the polygon yields a triangulation of the input n-point set. A nontrivial upper bound on the worst-case performance of the heuristic in terms of n and another parameter is derived. Under the assumption of uniform point distribution it is observed that the heuristic yields a solution within the factor of $O(\log n)$ from the optimum almost certainly, and the expected length of the resulting triangulation is of the same order as that of a minimum length triangulation. The heuristic runs in time $O(n^3)$.

Key words. planar point set, minimum weight triangulation, minimum spanning tree, heuristic, polygon, running time, uniform point distribution, almost certainly

AMS(MOS) subject classification. 68C05

1. Introduction. Let S be a set of n points in the plane. A *triangulation* of S is a maximal set of nonintersecting straight-line segments between these points. Any triangulation of S partitions the *convex hull* of S (see [6], [16]) into triangles. For a set of straight-line segments in the plane, T, let $|T|$ denote the total length of the segments in T. A *minimum weight triangulation* is any triangulation which minimizes $|T|$ among all triangulations T of S. Minimum weight triangulations have an application in interpolating values of two-argument functions [14], [16].

There are two well-known triangulation algorithms: the greedy triangulation and the Delaunay triangulation (see [4], [10], [12]–[16]). The former inserts a segment into the plane if it is the smallest among all segments between points in S not intersecting those already in the plane. The latter simply constructs the dual of the Voronoi diagram for S. Manacher and Zobrist [15] show that neither approximates the optimum. Let $GT(S)$ and $DT(S)$ denote the outcome of the greedy triangulation and the Delaunay triangulation of S, and let $M(S)$ stand for the total length of a minimum weight triangulation of S. Specifically, for arbitrarily large n Manacher and Zobrist construct sets of n points in the plane, S', S'', such that

$$|GT(S')|/M(S') = \Omega(n^{1/3}),$$

$$|DT(S'')|/M(S'') = \Omega(n/\log n).$$

In [10], Levcopoulos improves the former result to $\Omega(\sqrt{n})$ by constructing another set family, and in [8], Kirkpatrick strengthens the latter result by exhibiting sets S''' for which

$$|DT(S''')|/M(S''') = \Omega(n).$$

In this paper, we propose a new triangulation algorithm. Its idea is simple. First, we find the convex hull of S. Next, we construct a minimum length planar forest connecting the convex hull with the remaining points in S. The convex hull plus the forest result in a polygon (we may assume that the edges of the forest are doubled). Finally, a minimum weight triangulation of the polygon is found. The union of the polygon triangulation with the polygon yields a triangulation of S. The entire algorithm runs in time $O(n^3)$.

* Received by the editors September 7, 1984; accepted for publication (in revised form) April 21, 1987.

† Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden.

For example, the result of the new triangulation run on the set S' for which the greedy triangulation does not behave well is not worse than that of the triangulation BT of S' constructed in Lemma 5 in [15], since BT trivially includes all edges of the polygon constructed by the new triangulation. Hence, by Lemma 5 in [15], the result of the new triangulation of S' is $\Omega(n^{1/3})$ times better than that of the greedy triangulation of S'. The result of the new triangulation run on the set S'' for which the Delaunay triangulation is so bad is trivially optimal since all points in S' lie on the convex hull of S' (see [8]).

Let jump (S) be the smallest real d such that starting from any point in S, we can reach the convex hull by jumping from one point in S to another one etc., provided that the length of no jump is greater than d. Let $NT(S)$ denote the new triangulation of S. The optimality of the forest connecting S with the convex hull enables us to prove

$$|NT(S)| = O(\log n \times M(S) + n \times \text{jump } (S)).$$

Thus, $|NT(S)|$ is within the factor of $O(\log n + n \times \text{jump } (S)/M(S))$ from the optimum. If jump (S) is small enough, the approximation is satisfactory. It is easy to find, for arbitrarily large n, sets of n points in the plane, S, where $n \times \text{jump } (S)/M(S)$ can be as bad as $\Omega(n)$. One example would be a set of n points, $n - 2$ placed on an arc of length $1/n$ of a circle of radius 1, one in the center of the circle, one on the perimeter of the circle on the opposite side of the arc. However, sometimes it is sufficient to expand S by relatively few new points in order to decrease jump (S) dramatically. Note that for any $\alpha > 0$ with $0 < \alpha < 1$, if jump (S) is not greater than $O(n^{-\alpha})$ times the length of the convex hull of S, then

$$|NT(S)|/M(S) = O(n^{1-\alpha}).$$

If we assume that the n points of S are uniformly distributed in a given square, then we have $|NT(S)|/M(S) = O(\log n)$ *almost certainly*, i.e., with the probability of at least $1 - cn^{-\alpha}$, where c, α are constants satisfying $c > 0$, $\alpha > 1$. Also, the expected length of $|NT(S)|$ is of the same order as that of minimum triangulation of S, then. In [12] and [13], under the same assumptions, Lingas has shown the greedy and the Delaunay triangulation to be within a logarithmic factor of the optimum, almost certainly. On the other hand, Chang and Lee [3] have shown that the expected length of the Delaunay triangulation is of the same order as that of a minimum weight triangulation, under the uniform point distribution.

2. The new triangulation.

Specification. To define the new triangulation we shall use the following conventions:

(a) S stands for a set of n real-coordinate points in the plane.

(b) The set of $n(n - 1)/2$ straight-line segments whose ends are in S is denoted by $E(S)$. Elements of $E(S)$ are called edges.

(c) The set of edges on the convex hull of S is denoted by $CH(S)$.

(d) The figure composed of $CH(S)$ and the points of S lying inside $CH(S)$ is denoted by $C(S)$.

(e) A *polygon* means a sequence of closed straight-line segments $s_1, s_2, \cdots, s_k$, such that if $\#(s_i \cap s_j) > 1$ then $s_i = s_j$ for $1 \leqq i, j \leqq k$, and the union of the segments partitions the plane into two disjoint, connected regions. The finite region forms the *inside* of the polygon. The endpoints of the above segments have real coordinates and are called *vertices* of the polygon. Given a polygon P, there always exists a sequence $(a_0, \cdots, a_m)$ of not necessarily distinct vertices of P such that $P = \{[a_m, a_0]\} \cup \{[a_i, a_{i+1}] | 0 \leqq i < m\}$.

(f) A *spanning forest* of $C(S)$, F, is any collection of nonintersecting edges in $E(S)$ such that $CH(S) \cup F$ is a polygon and each point in S is an endpoint of an edge in $CH(S)$ or F. Note that according to (e), the polygon $CH(S) \cup F$ can be represented by a sequence of its vertices including each edge in F twice as a subsequence. A *minimum spanning forest* of $C(S)$ is a spanning forest of F achieving the smallest possible total edge length.

The above conventions will be used throughout the entire paper. Employing them, we define a general triangulation algorithm as follows.

ALGORITHM 1.

(1) Find the convex hull of S, $CH(S)$;

(2) Find a minimum spanning forest of $C(S)$, F;

(3) Find a minimum weight triangulation of the polygon $CH(S) \cup F$;

(4) Output the union of the triangulation of $CH(S) \cup F$ and $CH(S) \cup F$ as the triangulation of S.

Implementation. The convex hull of S can be constructed in time $O(n \log n)$ (see [6], [16]). To construct the forest F, we can use an algorithm analogous to Shamos's algorithm for Euclidean minimum spanning tree in the plane [16].

LEMMA 1. *A minimum spanning forest of $C(S)$ can be constructed in time $O(n \log n)$.*

Proof. Let $G(S)$ be the complete weighted graph on S, where edge weights are equal to the lengths of the corresponding edges in $E(S)$. Let us augment $G(S)$ by a virtual vertex connected by a zero-weight edge with each vertex on $CH(S)$, and not adjacent to any vertex inside $CH(S)$.

To see that the problem of constructing a minimum spanning forest of $C(S)$ reduces to that of constructing a minimum spanning tree of the augmented graph recall Prim's algorithm for minimum spanning tree [1], [7], [17]. Apply this algorithm to the augmented graph, beginning with the virtual vertex labeled and all vertices in S unlabeled. At each subsequent step, we add a shortest edge between a labeled and unlabeled vertex. Let k be the number of vertices in the augmented graph that are points in $S \cap CH(S)$. Clearly, after the first k steps, all vertices on $CH(S)$ become labeled and connected by zero-weight edges with the virtual vertex. Now, recall the definition of the Delaunay triangulation of S [16]. By Lemma 6.2 in [16], at each, next subsequent step of Prim's algorithm applied to the augmented graph, the added edge is an edge of the Delaunay triangulation of S incident to the labeled (nonvirtual) vertex. When all vertices become labeled, we obtain a minimum spanning tree of the augmented graph where all edges of the tree that are not incident to the virtual vertex are edges of the Delaunay triangulation of S. Hence, if we delete the virtual vertex from the minimum spanning tree of the augmented graph, we obtain a minimum spanning forest of $C(S)$.

To find a minimum spanning tree of the augmented graph, it is enough to run Prim's algorithm on its subgraph containing only these edges that are either incident to the virtual vertex or are edges of the Delaunay triangulation of S. Since the Delaunay triangulation of an n-point planar set can be constructed in time $O(n \log n)$ [16], the above subgraph can also be determined in time $O(n \log n)$. As the subgraph has only $O(n)$ edges, Prim's algorithm can be implemented to run on it in time $O(n \log n)$ (see [16, p. 77]). □

Further, we shall assume that the forest F computed in the second step of Algorithm 1 is constructed as in the proof of the above lemma.

Finally, we can find a minimum weight triangulation of the polygon $CH(S) \cup F$ in cubic time by the following fact, independently proved in [5], [8] and [12].

Fact 1. A minimum weight triangulation of a simple polygon with n vertices can be found in time $O(n^3)$.

In conclusion, the new triangulation can be computed in time $O(n^3)$.

Worst-case analysis. To derive the upper bound on the approximation factor of the new triangulation, we need to introduce the concept of jump (S) formally.

The smallest real d such that there is a spanning forest of $C(S)$ consisting only of edges of length not exceeding d is denoted by jump (S).

To note that every edge in a minimum spanning forest of $C(S)$ is of length not greater than jump (S), we prove the following simple lemma.

LEMMA 2. *Let T be a minimum weight spanning tree of a graph G. The tree T also minimizes the weight of the heaviest edge among all spanning trees of G.*

Proof. Let d be the minimum real such that there is a spanning tree of G with all edges of weight not exceeding d. Let $e_1, e_2, \cdots e_k$ be the edges of T in nondecreasing order. We shall prove by induction on j, $1 \leqq j \leqq k$, that for $1 \leqq i < j$, e_i are edges of a spanning tree U of G whose all edges are of weight $\leqq d$. Assume the inductive hypothesis for $j < k$. We have $|e_j| \leqq d$ since otherwise $|U| < |T|$. If e_j is in U then we are done. Otherwise, we add e_j to U and delete an edge of T on the cycle closed by e_j that is not in U to obtain a spanning tree of G satisfying the inductive hypothesis for $j + 1$. □

By the definition of the forest F of $C(S)$ in Algorithm 1, we obtain the following corollary from Lemma 2.

LEMMA 3. *Each edge of the forest F constructed in the second step of Algorithm* 1 *has length not exceeding* jump (S).

Proof. After extending F by the virtual vertex connected by zero-weight edges to all vertices of F on $CH(S)$, we obtain a minimum weight spanning tree of the augmented graph defined in the proof of Lemma 1. Now, the lemma follows from Lemma 2. □

The following technical definition and theorem lead to upper bounds on $|NT(S)|$ in terms of $M(S)$ and jump (S).

Consider a triangulation T of S and a spanning forest F of $C(S)$. For every edge e in F, we define $A(T, e)$ as the set of edges s in T such that:

> s crosses e and at least one of the two pieces of s between the crossing point and an endpoint of s is not crossed by other edges in F.

THEOREM 1. *Let T be a triangulation of S. Let F be a spanning forest of $C(S)$. There exists a triangulation of the polygon $CH(S) \cup F$ of edge length at most*

$$6 \sum_{e \in F} \#A(T,e) \times |e| + (3 \log n + 9)|T| + (3 \log n + 2)|F|.$$

Before proving Theorem 1, let us see how it induces the upper bound on $|NT(S)|/M(S)$ in terms of jump (S). Assume that in Theorem 1, T is a minimum weight triangulation of S, and F is a minimum spanning forest of $C(S)$. By the Euler formula for planar graphs [7], T as a triangulation of S has at most $3n - 6$ edges. Since every edge in T can occur in at most two different sets $A(T, e)$, we have

$$\sum_{e \in F} \#A(T,e) \leqq 2 \times (\#T) \leqq 6n - 12.$$

By Theorem 1, the minimum weight triangulation of $CH(S) \cup F$ that is a part of $NT(S)$ is of length no greater than

$$6 \sum_{e \in F} \#A(T,e)|e| + (3 \log n + 9)|T| + (3 \log n + 2)|F|$$
$$\leqq (36n - 72) \times \text{jump}(S) + (3 \log n + 9) \times M(S) + (3 \log n + 2)|F|.$$

Since any triangulation of S includes in particular a spanning forest of $C(S)$ and the convex hull of S, we have $|F| + |CH(S)| \leqq M(S)$. Putting everything together, we obtain the following.

COROLLARY 1. $|NT(S)| \leqq (6 \log n + 12) \times M(S) + 36n \times \text{jump}(S)$.

In particular, we have the following.

COROLLARY 2. *For any real* α *with* $0 < \alpha < 1$, *if* jump $(S) = O(|CH(S)|/n^{\alpha})$, *then*

$$|NT(S)|/M(S) = O(n^{1-\alpha}).$$

As the proof of Theorem 1 is quite involved, we precede it with the following introduction: At the beginning, the polygon $CH(S) \cup F$ is partitioned into smaller polygons. Next, specific triangulations of the smaller polygons are considered and upper bounds on their length are derived. The upper bounds are expressed in terms of the length of pieces of T and F. The union of the triangulations of the smaller polygons and the contours of the smaller polygons results in a triangulation of the polygon $CH(S) \cup F$. By summing the lengths of the component triangulations and contours, we obtain an upper bound on the length of a minimum weight triangulation of $CH(S) \cup F$, i.e., on $|NT(S)|$ in particular. The upper bound is the thesis of Theorem 1. Among the smaller polygons into which $CH(S) \cup F$ is partitioned, the important ones are polygons denoted by $P_i(e)$, where $e \in F$, $i = 1, 2$. The total length of the specific triangulations of the other smaller polygons can easily be expressed in terms of the total length of contours of $P_i(e)$'s, $|T|$ and $|F|$. It turns out that polygons $P_1(e)$ and $P_2(e)$ can be triangulated by drawing lines of the length proportional to the length of the pieces of T inside them plus $\#A(e, T)|e|$. As e may be much longer than particular edges in T crossing $P_1(e)$ or $P_2(e)$, the value of $\#A(e, T)|e|$ may considerably exceed the length of the pieces of T inside $P_1(e)$ and $P_2(e)$.

The formal proof of Theorem 1 is the remaining part of this section.

Let e be an edge in the forest F, and let H_1, H_2 be the two half-planes induced by the line colinear with e. For $i = 1, 2$, let $P_i(e)$ be the polygon $(q_0, q_1, \cdots, q_{k+1})$ (see Fig. 1) such that:

(a) q_0 and q_{k+1} are the endpoints of e;

(b) $q_1, \cdots, q_k$ lie in H_i;

(c) There are points $p_1, \cdots, p_k$ inside (q_0, q_{k+1}) such that for $j = 1, \cdots, k$, (q_j, p_j) is an initial segment of an edge $(q_j, -)$ in the triangulation T such that (q_j, p_j) lies within the polygon $(q_0, q_1, \cdots, q_{k+1})$ and is not crossed by any edge in F;

(d) The polygon $(q_0, q_1, \cdots, q_{k+1})$ is of the maximum number of vertices among all polygons satisfying (a)–(d).

Next, let Tr (F) be the set of all triangles in T whose three edges are not crossed inside by any edge in F (see also Fig. 1).

We have the following remark and lemma on $P_i(e)$'s and Tr (F).

Remark 1. If the inside of an edge (v_1, v_2) in T is crossed by an edge in F then for $j = 1, 2$, there exist a point $w \in [v_1, v_2]$ and a polygon $P_i(e)$ such that the segment (v_j, w) lies inside $P_i(e)$.

Proof. Let e be the edge of F that crosses the inside of (v_1, v_2) closest to v_j. Then, if w is the crossing point of e and (v_1, v_2), then (v_j, w) lies either inside $P_1(e)$ or inside $P_2(e)$. □

LEMMA 4. *The convex hull of* S *is partitioned into the polygons* $P_i(e)$, *triangles in* Tr (F), *and some inner convex polygons by drawing the contours of the polygons* $P_i(e)$, *the triangles in* Tr (F) *and the convex hull of* S.

Proof. The inside of each $P_i(e)$ is disjoint from S and the edges in F. Otherwise, a point p in S would lie inside some quadrilateral $(p_j, p_{j+1}, q_{j+1}, q_j)$, or some triangle

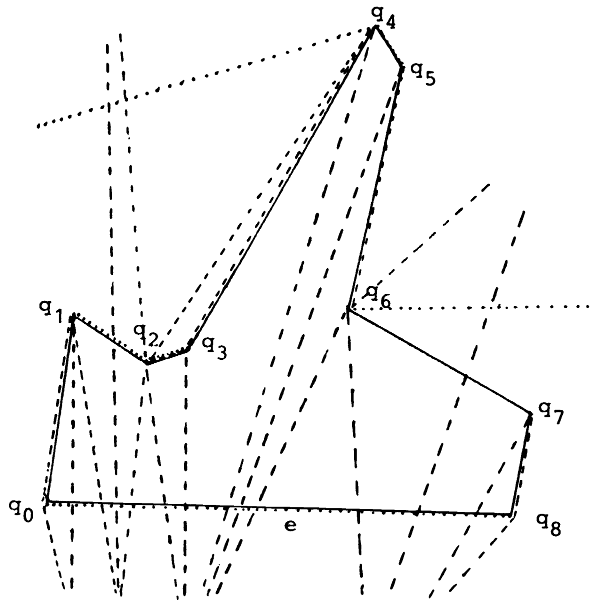


FIG. 1. *An example of a polygon $P_i(e)$. The edges of $P_i(e)$, T and F are respectively marked with continuous, broken and dotted lines. Notice that the triangle* (q_2, q_3, q_4) *is in* Tr (F).

(p_{j+1}, q_j, q_{j+1}) or (p_j, q_j, q_{j+1}), assuming the notation from the definition of $P_i(e)$. To consider only the quadrilateral case, we may assume that $p_j = q_j$ or $p_{j+1} = q_{j+1}$, respectively, in the triangular case. Further, we may assume without loss of generality that the point p is closest to e among all points in S inside $(p_j, p_{j+1}, q_{j+1}, q_j)$. Since the sides (q_j, p_j), (q_{j+1}, p_{j+1}) of the quadrilateral are fragments of edges of T, at least one edge d of T incident to p has to cross e between p_j and p_{j+1}. The fragment of d between p and e cannot be crossed by F since otherwise either there is a point in S inside $(p_j, p_{j+1}, q_{j+1}, q_j)$ closer to e than p is or F crosses (q_j, p_j) or (q_{j+1}, p_{j+1}) (we obtain a contradiction with the definition of p or $P_i(e)$, respectively). Hence, the point p could be added to the set of vertices of $P_i(e)$ without violating the conditions (a)–(c) from the definition of $P_i(e)$. Thus, $P_i(e)$ could not satisfy the maximality requirement in this case. We conclude that the inside of each $P_i(e)$ is disjoint from S and the edges in F.

On the other hand, by the definition of the polygons $P_i(e)$, no edge of a triangle in Tr (F) can intersect two edges of $P_i(e)$. Moreover, no polygon $P_i(e)$ is in Tr (F). Putting everything together, we conclude that the insides of $P_i(e)$'s and triangles in Tr (F) are pairwise disjoint (see Fig. 2).

Draw the perimeters of the polygons $P_i(e)$, the triangles in Tr (F) and the convex hull of S. Consider a polygonal face P in the resulting partition, different from $P_i(e)$'s and triangles in Tr (F), lying within $CH(S)$. Let an *initial fragment* of an edge e mean a segment of e at least one endpoint of which is also an endpoint of e.

First, we shall prove that no initial fragment of any edge in T that is incident to a vertex v of P can lie inside P. Suppose otherwise. Let d be an edge in T violating the above claim. Naturally, there is another edge d' in T incident to v such that d and d' induce a triangular face t in T.

If none of the edges of t is crossed inside by F then t is in Tr (F). Since, by the definition of d, t overlaps with P, we obtain a contradiction with the definition of P. If an edge f of F crosses the inside of an edge of t, then either it crosses the inside of another

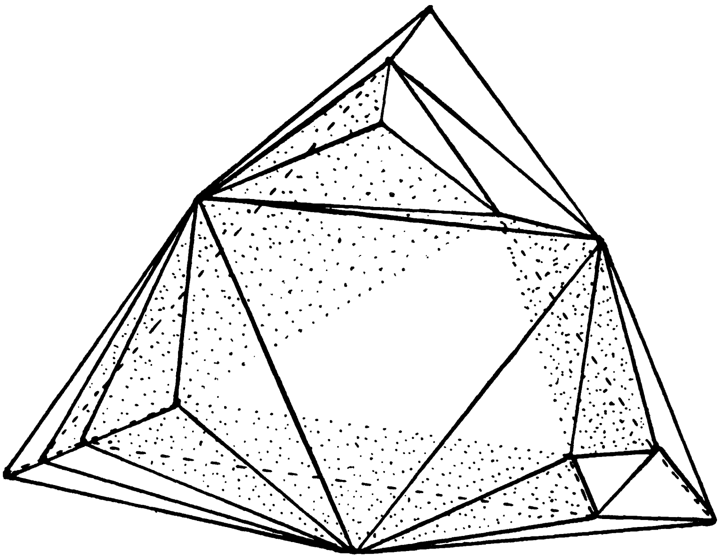


FIG. 2. *An example of partitioning the convex hull of S into polygons $P_i(e)$, triangles in* Tr (F), *and inner convex polygons. The edges of T are marked with continuous lines, and those of F with broken lines. The polygons $P_i(e)$ are darkened with points.*

edge of t or ends at the apex of t opposite to the crossed edge. If f crosses d then P overlaps with a polygon $P_i(e)$ by Remark 1 and again we obtain a contradiction. Thus, we may assume without loss of generality that either f crosses the two remaining edges of t or f crosses one of the remaining edges and is incident to an endpoint of d. We may also assume without loss of generality that no other edge of F crosses t between f and d. In either case, the piece of t cut off by f and including d lies within one of the polygons $P_1(f)$, $P_2(f)$. Again we obtain a contradiction. In this way, we have proved that no initial fragment of an edge in T incident to a vertex of P lies inside P.

Let q be a vertex of P, and let p, r be the vertices of P incident to q. Suppose that the angle (p, q, r) in the inside of P is of more than 180 degrees. Then, an initial fragment of an edge in T incident to q must lie inside of P, which yields a contradiction. Thus P is a convex polygon. □

By Lemma 4, the following four propositions imply Theorem 1.

PROPOSITIONS.

(A) *For $e \in F$, $i \in \{1, 2\}$, there exist triangulations $T_i(e)$ of $P_i(e)$'s such that*

$$\sum_{e\in F}\sum_{i=1}^{2} T_i(e) \leqq 6\times \sum_{e\in F} \#A(T,e)\times|e|+6\times|T|.$$

(B) *The total length of the union of the contours of the polygons $P_i(e)$ is not greater than* $2|T| + 3|F|$.

(C) *The total length of the union of the contours of the triangles in* Tr (F) *and the convex hull of S is at most* $|T|$.

(D) *There exist triangulations of the inner convex polygons of total length not greater than* $3\lfloor\log n\rfloor|T| + 3\lfloor\log n\rfloor|F|$.

To prove Proposition A, we use the three following definitions.

(a) A polygon $P = (q_0, \cdots, q_{k+1})$ is *moderately visible* from its boundary edge (q_0, q_{k+1}) if for all vertices q_j, $1 \leqq j \leqq k$, there are points p_j inside (q_0, q_{k+1}) such that the segments (q_j, p_j) lie inside P (i.e., q_j can see p_j within P), and for any j', where $1 \leqq j' \leqq k$ and $j \neq j'$, the open segments (q_j, p_j), $(q_{j'}, p_{j'})$ do not intersect.

(b) Given a polygon $P = (q_0, \cdots, q_{k+1})$, for $j = 0, \cdots, k+1$, $\mathrm{dis}_P(j)$ denotes the minimum distance from q_j to a point in $[q_0, q_{k+1}]$. Next, given a triangulation U of P, for $j = 0, \cdots, k+1$, $n_U(j)$ is the number of edges (q_l, q_j) in U such that $\mathrm{dis}_P(l) \leqq \mathrm{dis}_P(j)$.

(c) Given three points in the plane, a, b, c, (a, b, c) stands for the angle that results from counterclockwise turning a half-line anchored at b, from the line induced by a and b to that induced by b and c.

Note that for any $e \in F$ and $i \in \{1, 2\}$, the polygon $P_i(e)$ is moderately visible from e. Hence, it will turn out that the following lemma provides a satisfactory candidate for the triangulations $T_i(e)$.

LEMMA 5. *Let $P = (q_0, \cdots, q_{k+1})$ be a polygon moderately visible from its boundary edge (q_0, q_{k+1}). There is a triangulation U of P, such that for $j = 1, \cdots, k$, $n_U(j) \leqq 3$.*

Proof. First, we shall partition P into convex polygons by drawing a set D of diagonals lying within P. Then, we shall triangulate the resulting convex polygons to obtain a complete triangulation of P.

To produce the convex partition D of P, we proceed as follows. First, observe the following fact:

For $j = 1, \cdots, k$, there is a unique vertex $q_{r(j)}$, $j < r(j) \leqq k+1$, such that $\mathrm{dis}_P(j) \geqq \mathrm{dis}_P(r(j))$, and for any vertex $q_{r'}, j < r' < r(j)$, it holds $\mathrm{dis}_P(r') > \mathrm{dis}_P(j)$. Analogously, for $j = 1, \cdots, k$, there is a unique vertex $q_{l(j)}$, $0 \leqq l(j) < j$, such that $\mathrm{dis}_P(j) \geqq \mathrm{dis}_P(l(j))$, and for any other vertex $q_{l'}$, $l(j) < l' < j$, it holds $\mathrm{dis}_P(l') > \mathrm{dis}_P(j)$.

Note that for $j = 1, \cdots, k$, q_j can see $q_{r(j)}$ or $r(j) = j+1$. Otherwise, there would exist a vertex $q_{r'}$ such that $j > r' > r(j)$ and $(q_j, q_{r(j)})$ crossed $(q_{r'-1}, q_{r'})$. Let Re (j) be the region of the points in the plane that are within the distance $\mathrm{dis}_P(j)$ from e. Clearly, the region Re (j) is convex. Since q_j and $q_{r(j)}$ can see at least one point on e within P and Re (j), the point $q_{r'}$ had to lie within Re (j). Hence, we have $\mathrm{dis}_P(r') \leqq \mathrm{dis}_P(j)$ which yields a contradiction with the definition of $r(j)$ (see Fig. 3). Analogously, for $j = 1, \cdots, k$, q_j can see $q_{l(j)}$ or $l(j) = j-1$. Clearly, if q_j is a concave vertex of $P_i(e)$ then $1 \leqq j \leqq k$. Let D be the set of diagonals $(q_j, q_{r(j)})$, where q_j is a concave vertex of P, $r(j) > j+1$ and $1 \leqq j \leqq k-1$, and the diagonals $(q_{l(j)}, q_j)$, where q_j is a concave vertex of P, $l(j) < j-1$ and $2 \leqq j \leqq k$. By the definition of D, any vertex q_j of $P_i(e)$ is incident to at most two diagonals in D whose other endpoints are within the distance from e not greater than that from e to q_j.

We claim that no two diagonals in D properly intersect each other. Suppose otherwise. First, consider the case where the intersecting diagonals are of the form $(q_j, q_{r(j)})$, $(q_{l(j')}, q_{j'})$ where $1 \leqq j \leqq k-1$ and $2 \leqq j' \leqq k$. We may assume without loss of gener-

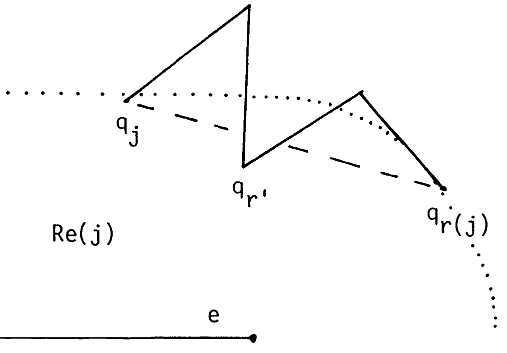


FIG. 3. *The dotted lines mark the boundary of* Re (j).

ality $\mathrm{dis}_P\,(j) \leqq \mathrm{dis}_P\,(j')$. Since $l(j') < j < j'$, we obtain a contradiction with the definition of $l(j')$. In turn, consider the case where the intersecting diagonals are in the form either $(q_j, q_{r(j)})$, $(q_{j'}, q_{r(j')})$, where $1 \leqq j < j' \leqq k - 1$ or $(q_{l(j)}, q_j)$, $(q_{l(j')}, q_{j'})$, where $2 \leqq j < j' \leqq k$. We may assume without loss of generality the first possibility. By $j < j' < r(j)$, we have $\mathrm{dis}_P\,(j') > \mathrm{dis}_P\,(j)$. Since $j' < r(j) < r(j')$ and $\mathrm{dis}_P\,(j) \geqq \mathrm{dis}_P\,(r(j))$, we obtain a contradiction with the definition of $r(j')$.

Thus, the set D forms a diagonal partition of P. Next, we claim that D is a convex partition under the assumption that angles of 180 degrees are considered as convex.

The proof of the last claim is as follows. Consider the diagonals of D incident to a concave vertex q_j of P that are in the form $(q_j, q_{r(j)})$ or $(q_j, q_{l(j)})$. They partition the inner angle at q_j into at most three angles. We can classify the resulting angles into two categories. The angles of the first category are in the form (q_l, q_j, q_r) where $\mathrm{dis}_P\,(l) \leqq \mathrm{dis}_P\,(j)$ and $\mathrm{dis}_P\,(r) \leqq \mathrm{dis}_P\,(j)$. Since the region Re (j) is convex, the angles of the first category are of no more than 180 degrees. The angles of the second category are in the form $(q_{r(j)}, q_j, q_{j+1})$ or $(q_{j-1}, q_j, q_{l(j)})$. By symmetry, it is sufficient to show that the former angle is of less than 180 degrees. We argue as follows. Since q_{j+1} can see a point on e, it can see also a point inside $(q_j, q_{r(j)})$. Hence, the above angle is of less than 180 degrees. We conclude that the set D is a convex partition of P. Importantly, for $j = 1, \cdots, k$, there are at most two diagonals in D of the form (q_i, q_j) where $\mathrm{dis}_P\,(i) \leqq \mathrm{dis}_P\,(j)$ by the definition of D.

To complete D to a full triangulation of P, consider a convex face C in the partition of P induced by D. Let $b(C)$ be the edge of C through which all vertices of P that are also vertices of C and are not endpoints of $b(C)$ see some points on e within P. By convention, if C is bounded by e then $b(C) = e$. Next, let $v(C)$ be an endpoint of $b(C)$ that is closest to e. By arguing as for the angles of the second category, we observe that no other edge of C can be co-linear with $b(C)$. Therefore, if we connect each vertex of C not adjacent to $v(C)$ with $v(C)$ by a diagonal then we obtain a triangulation $T(C)$ of C. On the other hand, given a vertex q_j of C different from the endpoints of $b(C)$, at least one endpoint of $b(C)$ is in Re (j) since otherwise q_j could not see any point on e through $b(C)$. Hence, by the definition of $v(C)$ and $T(C)$, for each vertex q_j of C, there is at most one diagonal in $T(C)$ of the form (q_i, q_j), where $\mathrm{dis}_P\,(i) \leqq \mathrm{dis}_P\,(j)$.

We conclude that the union U of D and the triangulations $T(C)$ is a triangulation of P satisfying $n_U(i) \leqq 3$ for $i = 1, \cdots, k$. □

Let $T_i(e)$ be a triangulation of the polygon $P = P_i(e)$ satisfying the thesis of Lemma 5. Given an edge $d = (p, q)$ in $T_i(e)$, let

$$m(d) = if\, \mathrm{dis}_P(p) \geqq \mathrm{dis}_P(q) \quad then\, p \quad else\, q.$$

Let us call a vertex of $P_i(e)$ *sound* if it is not an endpoint of e. For a sound vertex of $P_i(e)$, v, let $l(v)$ be the length of the longest edge in T that ends at v and crosses e. By these definitions and triangle inequalities, for any edge d in $T_i(e)$, we have $|d| \leqq |e| + 2l(m(d))$. Let $|\mathrm{Inside}\,(P_i(e)) \cap T|$ mean the total edge length of the piece of T inside $P_i(e)$. By Lemma 5, we have

$$\begin{aligned} |T_i(e)| &\leqq \sum_{d \in T_i(e)} |e| + 2l(m(d)) \\ &\leqq 3 \times \sum_{v \text{ is a sound vertex of } P_i(e)} |e| + 2l(v) \\ &\leqq \left(3 \times \sum_{v \text{ is a sound vertex of } P_i(e)} |e|\right) + 6 \times |\mathrm{Inside}\,(P_i(e)) \cap T|. \end{aligned}$$

By definition of $A(e, T)$ and Lemma 4, we have

$$\sum_{e\in F}\sum_{i=1}^{2}|T_i(e)| \leqq \sum_{e\in F}\sum_{i=1}^{2}\left(3\times \sum_{v \text{ is a sound vertex of } P_i(e)} |e|\right) + \sum_{e\in F}\sum_{i=1}^{2} 6\times|\text{Inside}\,(P_i(e))\cap T|$$

$$\leqq 6\times \sum_{e\in F} \#A(e, T))|e| + 6|T|$$

So, we have proved Proposition A.

Given a polygon P, we shall denote the contour of P by Contour (P). By triangle inequalities, the length of the contour of $P_i(e)$, i.e., $|\text{Contour}\,(P_i(e))|$, is not greater than $2|e| + 2|\text{Inside}\,(P_i(e)) \cap T|$. Hence, we have (B):

$$\left|\bigcup_{e\in F}\bigcup_{i=1}^{2}\text{Contour}\,(P_i(e))\right| \leqq \sum_{e\in F}\sum_{i=1}^{2}|\text{Contour}\,(P_i(e))| - |F|$$

$$\leqq \sum_{e\in F}\sum_{i=1}^{2} 2|e| + \sum_{e\in F}\sum_{i=1}^{2} 2|\text{Inside}\,(P_i(e))\cap T| - |F|$$

$$\leqq 3|F| + 2|T|$$

Proposition C is trivial. To prove Proposition D, we use the following lemma.

LEMMA 6. *Let P be a convex polygon with l vertices. A triangulation of P of the edge length* $\lfloor \log l \rfloor |\text{Contour}\,(P)|$ *can be constructed in time* $O(l)$.

Proof. To construct a triangulation of P of length $\leqq \lfloor \log l \rfloor |\text{Contour}\,(P)|$, we follow the perimeter of P, and then, the perimeter of the resulting, current convex subpolygon of P counterclockwise, connecting consecutive, nonadjacent vertices by diagonals (see Fig. 4). More formally, the triangulation procedure is as follows:

```
input: a list $L$ of vertices of $P$ in counterclockwise order;
output: a list $U$ of all edges of a triangulation of $P$;
$U \leftarrow$ empty list;
until $\#L < 4$ do
    begin
        for $i \leftarrow 1, 3$ do front $(i) \leftarrow$ the $i$th vertex on $L$;
        append the diagonal (front (1), front (3)) to $U$;
        move front (1) from the front to the end of $L$;
        delete front (2) from $L$
end
```

Let $p_0, \cdots, p_{l-1}$ be the input vertex sequence L. First, suppose that $l = 2^m$ for some natural number m. Note that the sequence of consecutive valuations of the variable index defined by front $(1) = p_{\text{index}}$ can be decomposed into maximal monotone subsequences $\alpha_0, \alpha_1, \cdots, \alpha_t$ of $\{0, 1, \cdots, l-1\}$ where $\#\alpha_0 = m/2$, $\#\alpha_{k+1} = \#\alpha_k/2$ and $\#\alpha_{k+1} \geqq 2$ for $k = 0, 1, 2, \cdots, t$. It follows that $t \leqq m - 2$. Each of the monotone subsequences corresponds to a closed chain of diagonals of P appended to U, forming a subpolygon of P. Hence, the total length of the diagonals in U does not exceed $(m-1)|\text{Contour}\,(P)|$ in this special case. In the general case, we have $l = 2^m + r$ where $m = \lfloor l \rfloor$. Since $r < 2^m$, the first r diagonals drawn by the triangulation procedure form a subpolygon of P. The subpolygon has exactly 2^m vertices. Hence, the total

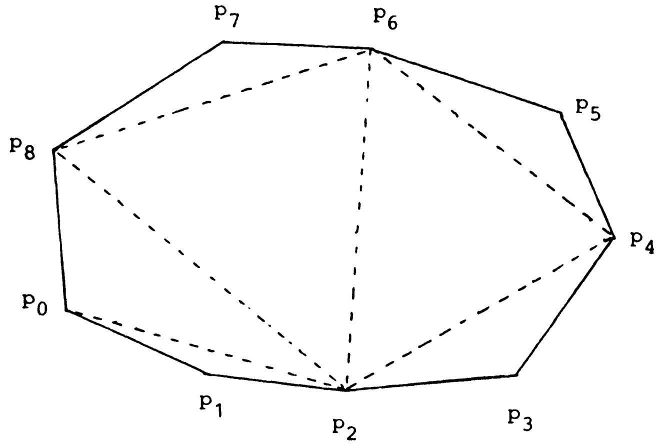


FIG. 4

length of the remaining diagonals in U is $(m - 1)|\text{Contour } (P)|$ by the previous case. We conclude that $|U| \leqq m|\text{Contour } (P)|$.

Since during the execution of the block under the until statement, one vertex is always deleted from L, the procedure runs in time $O(l)$. □

By Lemma 4 and Propositions B and C, the total length of the perimeters of the inner convex polygons resulting from drawing the perimeters of the polygons $P_i(e)$ and the triangles in Tr (F), and the convex hull of S does not exceed $3|F| + 3|T|$. Hence, by Lemma 6, there exist triangulations of the inner convex polygons of total length not exceeding

$$3\lfloor \log n \rfloor |F| + 3\lfloor \log n \rfloor |T|,$$

i.e., Proposition D holds.

Let V be the union of all triangulations $T_i(e)$, all contours of the polygons $P_i(e)$ and the triangles in Tr (F), and minimum weight triangulations of the inner convex polygons, minus the edges in F. By Lemma 4, V is a triangulation of $CH(S) \cup F$. By Propositions A–D, we have

$$|V| \leqq 6 \times \sum_{e \in F} \#A(T, e)|e| + (3 \log n + 9)|T| + (3 \log n + 2)|F|.$$

Probabilistic analysis. Let us assume the sample area from which the n points in S are drawn to be a unit square. It seems natural to assume the uniform point distribution, i.e., if B is a subset of the set of all points lying within the square, then for $i = 1, \cdots, n$, the probability that the ith point is in B is equal to the area of B. Following Angluin and Valiant [2], let *almost certainly* mean with the probability of at least $1 - cn^{-\alpha}$, where c, α are constants satisfying $c > 0$, $\alpha > 1$.

In [12], [13], Lingas among others showed that the length of the Delaunay triangulation of a point set S uniformly distributed in a unit square is within a logarithmic factor from $M(S)$ almost certainly. Following [13], we can partition the unit square into square cells of $O(\sqrt{\log n/n})$ width. Then, arguing as in [13, p. 26], we can prove that under the assumption of uniform point distribution all cells contain a point in S almost certainly. This implies $M(S) = \Omega(\sqrt{n/\log n})$ almost certainly (see [13]) and jump (S) =

$O(\sqrt{\log n/n})$ almost certainly. By Corollary 1, we can conclude that $|NT(S)|/M(S) = O(\log n)$ almost certainly.

In [3], Chang and Lee strengthened Lingas's result on Delaunay triangulation in the average case. They showed that the expected length of the Delaunay triangulation of a point set S uniformly distributed in a unit square is within a constant factor from the expected length of $M(S)$. By combining Corollary 1 with the technique of Chang and Lee, we could also prove that an analogous result holds for the new triangulation. However, there exists a shorter way of deriving the two probabilistic results for the new triangulation by using the following observation due to Christos Levcopoulos [11]: Since the minimum spanning forest F of $C(S)$ constructed in the proof of Lemma 1 is a subset of the Delaunay triangulation of S, the length of the new triangulation is never greater than that of the Delaunay triangulation of S.

By combining the above observation with Corollary 2.5 in [13], and Theorem 4.1 in [3], we have the following.

THEOREM 2. *Let S be a random set of n points which are uniformly distributed in a unit square.*

(1) *For any positive real $\alpha > 1$, we have*

$$Pr[|NT(S)|/M(S) = O(\alpha \times \log n)] \geqq 1 - n^{1-\alpha}/\log n.$$

(2) *Let $E(|NT(S)|)$ and $E(M(S))$ be the expected total length of the new triangulation and minimum weight triangulation, respectively. Then*

$$\frac{E(|NT(S)|)}{E(M(S))} = O(1).$$

Acknowledgments. I would like to express my great appreciation to Christos Levcopoulos, Joseph O'Rourke and an unknown referee for constructive suggestions and comments. Among others, Christos Levcopoulos has suggested Lemma 2, which has simplified the new triangulation, and observed that the new triangulation is never worse than the Delaunay triangulation.

REFERENCES

[1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] D. ANGLUIN AND L. G. VALIANT, *Fast probabilistic algorithms for Hamiltonian circuits and matchings*, Proc. of the 9th Annual ACM Symp. on Theory of Computing, Association for Computing Machinery, New York, 1978.

[3] R. C. CHANG AND R. C. T. LEE, *On the average length of Delaunay triangulations*, BIT, 24 (1984), pp. 269–273.

[4] R. D. DUPPE AND H. H. GOTTSCHALK, *Automatische Interpolation von Isolinen bei willkurlich stutzpunkten*, Allgemeine Vermessungsnachrichten, 77 (1970), pp. 423–426.

[5] P. D. GILBERT, *New results in planar triangulations*, M. S. thesis, Coordinated Science Laboratory, Univ. of Illinois, Urbana, IL, 1979.

[6] R. L. GRAHAM, *An efficient algorithm for determining the convex hull of a finite planar set*, Inform. Process. Lett., 1 (1972), pp. 132–133.

[7] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[8] D. G. KIRKPATRICK, *A note on Delaunay and optimal triangulations*, Inform. Process. Lett., 10 (1980), pp. 127–131.

[9] G. T. KLINCSEK, *Minimal triangulations of polygonal domains*, Ann. Discrete Math., 9 (1980), pp. 121–123.

[10] C. LEVCOPOULOS, *An $\Omega(\sqrt{n})$ lower bound for the nonoptimality of the greedy triangulation*, Inform. Process. Lett., 25 (1987), pp. 247–251.

[11] ———, *personal communication*, 1987.

[12] A. LINGAS, *Advances in minimum weight triangulation*, Ph.D. dissertation, Linköping Univ., Linköping, Sweden, 1983.

[13] ———, *The greedy and Delaunay triangulations are not bad in the average case*, Inform. Process. Lett., 22 (1986), pp. 25–31.

[14] E. L. LLOYD, *On triangulations of a set of points in the plane*, Proc. of the 18th Annual IEEE Conference on the Foundations of Computer Science, Providence, RI, 1977.

[15] G. K. MANACHER AND A. L. ZOBRIST, *Neither the greedy nor the Delaunay triangulation of a planar point set approximates the optimal triangulation*, Inform. Process. Lett., 9 (1979), pp. 31–34.

[16] F. P. PREPARATA AND M. I. SHAMOS, *Computational Geometry, An Introduction*, Texts and Monographs in Computer Science, Springer-Verlag, New York-Berlin, 1985.

[17] R. E. TARJAN, *Data Structures and Network Algorithms*, CBMS-NSF Regional Conference Series in Applied Mathematics 44, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

ON MINIMUM CRITICALLY *n*-EDGE-CONNECTED GRAPHS*

MARGARET B. COZZENS† AND SHU-SHIH Y. WU†‡

Abstract. Let n be an integer with $n \geqq 2$. A graph G is called critically n-edge-connected if the edge-connectivity $\lambda(G) = n$ and for any vertex v of G, $\lambda(G - v) = n - 1$. The sizes of critically n-edge-connected graphs are important and interesting in applications in communication networks. The maximum graphs with this property have been characterized [2]. In this paper, we first discuss some properties of minimum graphs, then show that the problem of finding a minimum critically n-edge-connected spanning subgraph of a given graph G is NP-complete.

Key words. graph theory, edge-connectivity $\lambda(G)$, connectivity $\kappa(G)$, NP-completeness

AMS(MOS) subject classification. 05C41

1. Introduction.

Let n be a fixed integer with $n \geqq 2$. A graph G shall be called *n-edge-connected* if the edge-connectivity $\lambda(G) = n$. A graph G is called *critically n-edge-connected* if G is n-edge-connected and for any vertex v of G, $\lambda(G - v) = n - 1$. A graph G is called *n-connected* if the vertex connectivity, $\kappa(G) = n$. A graph G is called *critically n-connected* if G is n-connected and for any vertex v in G, $\kappa(G - v) = n - 1$. A graph G is a *minimum* (*maximum*) critically n-edge-connected graph if no critically n-edge-connected graphs with the same number of vertices has fewer (more) edges than G.

In a communication network and circuit design, reliability is often determined by the connectivity and edge-connectivity of the corresponding graph. Therefore it is important to investigate, for fixed n, critically n-connected graphs ([3], [7]), and critically n-edge-connected graphs. We characterized the maximum graphs in a subset of critically n-edge-connected graphs, for each $n \geqq 2$ in [2]. Here we investigate the minimum critically n-edge-connected graphs.

We use $\{x\}$ to denote the least integer greater than or equal to x, and $[x]$ the greatest integer less than or equal to x.

2. An example of a minimum critically *n*-edge-connected graph.

For any fixed integers n, m, $m \geqq n + 1$, Harary [5] constructed classes of graphs $H_{n,m}$, that are minimum n-connected. These same graphs are minimum critically n-edge-connected graph with order m. $H_{n,m}$ is constructed as follows:

Case 1. n is even. Let $n = 2r$. Then $H_{2r,m}$ has vertices $0, 1, 2, 3, \cdots, m - 1$ and two vertices i and j are adjacent if $i - r \leqq j \leqq i + r$ (where addition is taken modulo m). $H_{4,8}$ is shown in Fig. 1.

Case 2. n is odd ($n > 1$), m is even. Let $n = 2r + 1$ ($r > 0$). Then $H_{2r+1,m}$ is constructed by first drawing $H_{2r,m}$, and then adding edges joining vertex i to vertex $i + m/2$ for $1 \leqq i < m/2$. $H_{5,8}$ is shown in Fig. 2.

Case 3. n is odd ($n > 1$), m is odd. Let $n = 2r + 1$ ($r > 0$). Then $H_{2r+1,m}$ is constructed by first drawing $H_{2r,m}$, and then adding edges $[0, (m - 1)/2]$ and $[0, (m + 1)/2]$, and $[i, i + (m + 1)/2]$ for $1 \leqq i < (m - 1)/2$. $H_{5,9}$ is shown in Fig. 3.

In Case 1 and Case 2, $\deg_{H_{n,m}}(i) = n$, for all $i \in V(H_{n,m})$ so that $|E(H_{n,m})| = \frac{1}{2} \sum_{i \in V(H_{n,m})} \deg_{H_{n,m}}(i) = \frac{1}{2}\, n \cdot m$.

* Received by the editors October 1, 1986; accepted for publication (in revised form) April 30, 1987.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

‡ Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China.

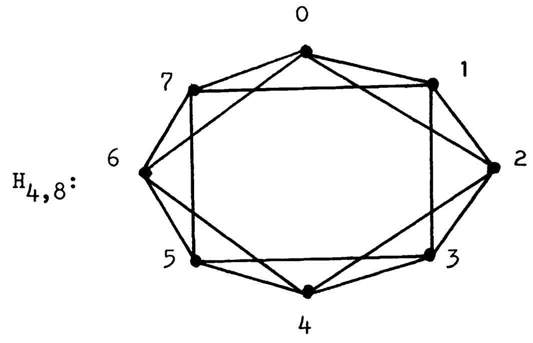


FIG. 1

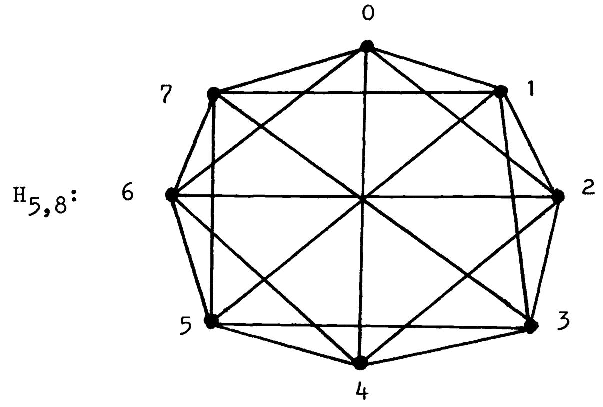


FIG. 2

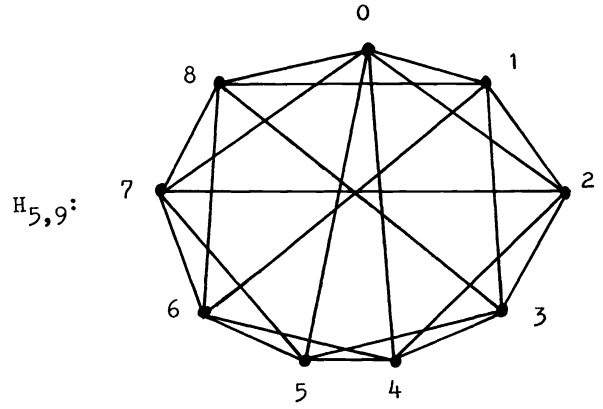


FIG. 3

In Case 3, $\deg_{H_{n,m}}(i) = n$, for $i = 1, 2, \cdots, m-1$, and $\deg_{H_{n,m}}(0) = n+1$. So that $\sum_{i\in V(H_{n,m})} \deg_{H_{n,m}}(i) = nm + 1 = 2\cdot|E|$. So $|E(H_{n,m})| = (nm+1)/2$.

Therefore for any fixed integers n, m, $m \geqq n+1$, $|E(H_{n,m})| = \{nm/2\}$.

Now we show that $H_{n,m}$ is a minimum critically n-edge-connected graph. $\delta(G)$ is the least degree over all vertices of G.

THEOREM 1. *The graph $H_{n,m}$ is n-connected* [5].

From the construction of $H_{n,m}$, it is clear that $\delta(H_{n,m}) = n$, and since

$$n \leqq \kappa(H_{n,m}) \leqq \lambda(H_{n,m}) \leqq \delta(H_{n,m}) = n,$$

we have $\lambda(H_{n,m}) = \delta(H_{n,m}) = n$. Therefore, we have the following theorem.

THEOREM 2. *The graph $H_{n,m}$ is n-edge-connected.*

For vertices j and k in a graph G, a (j, k)-*cutset* of G is a vertex cutset T such that j and k are in different components of $G - T$.

THEOREM 3. *The graph $H_{n,m} - \{i\}$ is $(n-1)$-connected, for any vertex i in $H_{n,m}$.*

Proof. Let $n = 2r$ if n is even, $2r+1$ if n is odd. The minimum degree, $\delta(H_{n,m} - \{i\})$, is $n-1$ so there exists a vertex cutset of size $n-1$. We will show that there is no vertex cutset with fewer than $n-1$ vertices.

Suppose there exists a vertex cutset T such that $2 \leqq |T| < n-1$. Let j and k be vertices belonging to different components of $(H_{n,m} - \{i\}) - T$ such that if i is between j and k then $0 \leqq k < i < j$, and if i is not between j and k then $j < k$. Define two vertex sets A and B in $H_{n,m} - \{i\}$ (addition is modulo m):

$$A = \{j, j+1, j+2, \cdots, k-1, k\},$$

$$B = \{k, k+1, k+2, \cdots, i-1, i+1, \cdots, j-1, j\}.$$

Note that $A \cup B = V(H_{n,m} - \{i\})$ and $A \cap B = \{j, k\}$. Since $|T| < n-1$, $|T| < 2r$. Therefore not both $T \cap A$ and $T \cap B$ can have r or more elements.

Case 1. $|T \cap A| < r$. $A - T = A - (A \cap T)$ so no more than $r-1$ consecutive elements are removed from A by T. Hence $A - T$ has a sequence of distinct vertices starting with j and ending with k with no difference greater than r between any pair of consecutive vertices. This sequence is a (j, k)-path in $(H_{n,m} - \{i\}) - T$, a contradiction to T being a (j, k)-cutset.

Case 2. $|T \cap B| < r$.

Subcase (i). $|T \cap B| < r-1$. As in Case 1, no more than $r-2$ consecutive elements are removed from B by T. Hence $B - T$ has a sequence of distinct vertices starting with k and ending with j, and the difference between any two consecutive vertices is at most $(r-1)+1 = r$. (There is an additional 1 in the gap between $i-1$ and $i+1$.) This sequence is a (k, j)-path of $(H_{n,m} - \{i\}) - T$, a contradiction to T being a (j, k)-cutset.

Subcase (ii). $|T \cap B| = r-1$. Since j and k are not in T,

$$|T \cap A| = |T| - |T \cap B| < n - 1 - (r-1) = n - r < r + 1.$$

If $|T \cap A| < r$ then Case 1 applies. Therefore $|T \cap A| = r$. $|A| + |B| = (m+2) - 1 = m+1$. Therefore not both of $|A|$ and $|B|$ can be greater than $\{(m+1)/2\}$, but at least one is greater than or equal to $\{(m+1)/2\}$.

Suppose $|A| \geqq \{(m+1)/2\}$. If there exists a sequence of vertices in $A - T$ beginning with j and ending with k such that no pair of consecutive terms has a difference $\geqq r+1$, then this sequence is a (j, k)-path in $(H_{n,m} - \{i\}) - T$, a contradiction to T being a (j, k)-cutset. Thus we may assume that every sequence of vertices in $A - T$ beginning with j and ending with k has a pair of consecutive terms with difference $\geqq r + 1$. In fact, since $|T \cap A| = r$, this difference is exactly $r + 1$, and there is only one such consecutive pair with difference $r + 1$. All other consecutive pairs have a difference of 1. Call the pair of vertices with difference $r + 1$, s and $s + r + 1$ in the sequence $A - T$. Thus we can write $A - T$ as $\{j, j + 1, j + 2, \cdots, s - 1, s, s + r + 1, \cdots, k - 1, k\}$. (Note that j can be s.) Split $A - T$ into two parts:

$$A_1 = \{j, j+1, \cdots, s-1, s\} \quad \text{and} \quad A_2 = \{s+r+1, s+r+2, \cdots, k-1, k\}.$$

The difference in consecutive terms in each A_i is 1, so there is an edge in $(H_{n,m} - \{i\}) - T$ between them. But $m \geqq n + 1 \geqq 2r + 1$ implies $m/2 \geqq r + \frac{1}{2} > r$ if m is even, and $(m + 1)/2 \geqq r + 1 > r$ if m is odd. Thus there are some $a_1 \in A$ and $a_2 \in A_2$ such that $a_2 = a_1 + [(m + 1)/2]$. The sequence $\{j, j + 1, \cdots, a_1, a_2, \cdots, k - 1, k\}$ is a (j, k)-path in $(H_{n,m} - \{i\}) - T$, a contradiction to T being a (j, k)-cutset.

If $|B| \geqq \{(m + 1)/2\}$ then the same argument applies since $n - 1 > |T| \geqq 2$ implies $n \geqq 4$, hence $r \geqq 2$, so there is an edge between $i - 1$ and $i + 1$ in $H_{n,m} - \{i\}$.

All that remains is to show that no vertex cutset of only one vertex exists for $H_{n,m} - \{i\}$. Suppose $T = \{p\}$ is a vertex cutset of $H_{n,m} - \{i\}$. Since $|T| < n - 1$, $n \geqq 3$.

Case 1. If $p = i - 1$ (equivalently $i = p + 1$), then $i + 1, i + 2, \cdots, m - 1, 0, \cdots, i - 2$ is a path containing all the vertices of $H_{n,m} - \{i, p\}$, a contradiction to T being a cutset of $H_{n,m} - \{i\}$.

Case 2. $p \neq i - 1$ and $p \neq i + 1$. Without loss of generality assume $i < p \leqq m - 1$. Now $P_1 = p + 1, p + 2, \cdots, m - 1, 0, 1, \cdots, i - 1$ is a path and $P_2 = i + 1, i + 2, \cdots, p - 1$ is a path in $(H_{n,m} - \{i\}) - \{p\}$. If n is even then $r \geqq 2$ and

$$\{i-1, i+1\} \in E(H_{n,m})$$

so there is only one component of $(H_{n,m} - \{i\}) - \{p\}$. If n is odd then there exists an edge between some x in P_1 and $x + [(m + 1)/2]$ in P_2, again contradicting $T = \{p\}$ being a cutset of $H_{n,m} - \{i\}$. Therefore, there exists no cutset with only one vertex, and the theorem is proved. QED

Since $n - 1 \leqq \kappa(H_{n,m} - \{i\}) \leqq \lambda(H_{n,m} - \{i\}) \leqq \delta(H_{n,m} - \{i\}) = n - 1$, we have $\lambda(H_{n,m} - \{i\}) = \delta(H_{n,m} - \{i\}) = n - 1$. Therefore, we have the following theorem.

THEOREM 4. *The graph $H_{n,m} - \{i\}$ is $(n - 1)$-edge-connected, for any vertex i in $H_{n,m}$.*

Now we can show the main theorem of this section.

THEOREM 5. *For any given positive integers m, n, $m \geqq n + 1$, there exists a minimum critically n-edge-connected graph with order m.*

Proof. By Theorem 2 and Theorem 4, $H_{n,m}$ is critically n-edge-connected. $|E(H_{n,m})| = \{nm/2\}$ and $|V(H_{n,m})| = m$.

Let $G = (V, E)$ be a critically n-edge-connected graph with $|V| = m$. Thus $\lambda(G) = n$, and for any vertex v in G, $\lambda(G) \leqq \delta(G) \leqq \deg_G v$. Hence

$$2 \cdot |E| = \sum_{v \in V(G)} \deg_G v \geqq m \cdot \delta(G) = m \cdot n.$$

So $|E| \geqq mn/2$. $|E|$ is an integer, hence $|E| \geqq \{mn/2\} = |E(H_{n,m})|$. So no critically n-edge-connected graph with m vertices has fewer edges than $H_{n,m}$. Therefore $H_{n,m}$ is a minimum critically n-edge-connected graph with order m. QED

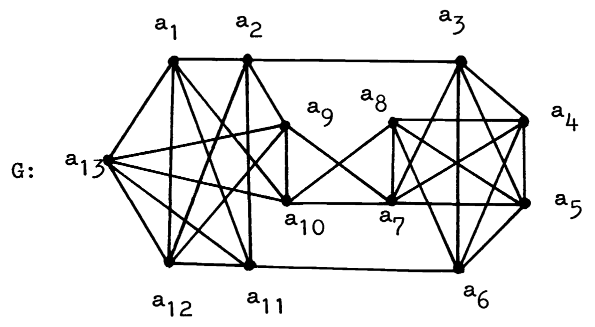


FIG. 4

3. Characterizations of minimum critically n-edge-connected graphs. In addition to $H_{n,m}$, there are other minimum critically n-edge-connected graphs. First we discuss some properties of minimum critically n-edge-connected graphs.

From the discussion of the graph $H_{n,m}$, it is easy to obtain the following lemma.

LEMMA 6. *If G is a minimum critically n-edge-connected graph with order m, then* $|E(G)| = \{mn/2\}$.

A graph G is called *almost regular of degree n* if there is at most one vertex of degree $n + 1$ and all other vertices have degree n. Clearly, an n-regular graph is almost regular of degree n.

THEOREM 7. *If $G = (V, E)$ is a minimum critically n-edge-connected graph, then G is almost regular of degree n. The proof follows from Lemma* 6.

The converse of Theorem 7 is not true. G, as shown in Fig. 4, is almost regular of degree 5, but G is not critically 5-edge-connected, since $\lambda(G) = 5$, and $\lambda(G - a_{10}) = 3 \neq 5 - 1$.

If G is n-edge-connected, then the order of G, m, is such that $m \geqq n + 1$. For $n + 1 \leqq m \leqq 2n$, we have a characterization of minimum critically n-edge-connected graphs.

THEOREM 8. *Let the order of G be m. For any n such that $n + 1 \leqq m \leqq 2n$, $G = (V, E)$ is a minimum critically n-edge-connected graph if and only if G is almost regular of degree n.*

To prove Theorem 8, we will use the following lemma.

LEMMA 9. *If G has m vertices and $\delta(G) \geqq [m/2]$, then $\lambda(G) = \delta(G)$* [1].

Proof of Theorem 8. By Theorem 7, if G is a minimum critically n-edge-connected graph, then G is almost regular of degree n.

Conversely, if G is almost regular of degree n, then $\delta(G) = n \geqq m/2 \geqq [m/2]$. By Lemma 9, we have $\lambda(G) = \delta(G) = n$. For any vertex $u \in V(G)$, $\delta(G - u) = n - 1 \geqq m/2 - 1$. Since $n - 1$ is an integer, $n - 1 \geqq \{m/2 - 1\}$.

Case 1. m is odd.

$$\delta(G-u)=n-1 \geqq \left\{\frac{m}{2}-1\right\}, \qquad n-1 \geqq \frac{m+1}{2}-1=\frac{m-1}{2}=\left[\frac{m-1}{2}\right].$$

Case 2. m is even.

$$\delta(G-u)=n-1 \geqq \left\{\frac{m}{2}-1\right\}=\frac{m-2}{2}=\left[\frac{m-1}{2}\right].$$

By Lemma 9, we have $\lambda(G - u) = \delta(G - u) = n - 1$.

$$|E(G)| = \frac{1}{2} \sum_{v \in V(G)} \deg_G v$$

$$= \begin{cases} \dfrac{mn}{2}, \text{ or} \\ \dfrac{1}{2}((m-1)n + n + 1) = \dfrac{1}{2}(mn+1) \end{cases}$$

$$= \left\{ \frac{mn}{2} \right\}.$$

Therefore, G is a minimum critically n-edge-connected graph. QED

The reader should note that G need not be n-connected in Theorem 8.

In general, the converse of Theorem 7 is not true, but if the vertex connectivity $\kappa(G) = n$, then we can give a characterization of minimum critically n-edge-connected graphs.

THEOREM 10. *Let $\kappa(G) = n$. $G = (V, E)$ is a minimum critically n-edge-connected graph if and only if G is almost regular of degree n.*

Proof. Let the order of G be m. By Theorem 7, we obtain the "only if part."

Conversely, if G is almost regular of degree n, then $\delta(G) = n$. Since $n = \kappa(G) \leqq \lambda(G) \leqq \delta(G) = n$, we have $\lambda(G) = n$.

For any vertex u in G, $\kappa(G - u) \leqq \lambda(G - u) \leqq \delta(G - u) = n - 1$. Suppose that $\lambda(G - u) < \delta(G - u)$, for some vertex u in G, then $\kappa(G - u) \leqq \lambda(G - u) < n - 1$. Thus, the connectivity $\kappa(G) < n$, a contradiction. So for any vertex u in G, we have $\lambda(G - u) = \delta(G - u) = n - 1$.

G is almost regular of degree n, so by the proof of Theorem 8, $|E(G)| = \{mn/2\}$.

Therefore, G is a minimum critically n-edge-connected graph. QED

The condition $\kappa(G) = n$ in Theorem 10 is necessary, since we can find a graph G, the one shown in Fig. 4, which is almost regular of degree n with $\kappa(G) < n$, G is a minimum n-edge-connected graph, but G is not critical with respect to $\lambda(G)$. Here $\kappa(G) = 4$, since $\{a_2, a_9, a_{10}, a_{11}\}$ is a vertex cutset.

For $m \geqq 2n + 1$, we can give some characterizations of minimum critically n-edge-connected graphs.

THEOREM 11. *For any given positive integers m, n, $m \geqq 2n + 1$, and $|V(G)| = m$, $G = (V, E)$ is a minimum critically n-edge-connected graph if and only if G is almost regular of degree n, and for each vertex u in a vertex cutset T with $|T| \leqq n - 1$, $\lambda(G - u) \geqq n - 1$.*

Proof. By Theorem 7, if G is a minimum critically n-edge-connected graph, then G is almost regular of degree n. Since G is critical with respect to $\lambda(G)$, for each vertex u in G, $\lambda(G - u) = n - 1$. So "the only if part" is complete.

Conversely, if G is almost regular of degree n, then $\delta(G) = n$. Since $\lambda(G - u) \geqq n - 1$ for some vertex u in G, and $\delta(G) = n$, we have $\lambda(G) \geqq n - 1$.

Suppose $\lambda(G) = n - 1$. Let E_1 be a minimum edge-cutset and G_1, G_2 be two components of $G - E_1$. $\delta(G) = n$ and $|E_1| = n - 1$, so $|V(G_1)| \geqq 2$ and $|V(G_2)| \geqq 2$. Since $m \geqq 2n + 1$, without loss of generality, we may let $|V(G_1)| \geqq n + 1$. Let A be the set of vertices in G_1 which are incident with E_1. $|A| \leqq n - 1$, since $|E_1| = n - 1$. So A is a vertex cutset with $|A| \leqq n - 1$, and for any vertex u in A, $\lambda(G - u) \leqq n - 2$, a contradiction.

Therefore $\lambda(G) > n - 1$. $n - 1 < \lambda(G) \leqq \delta(G) = n$, so $\lambda(G) = n$. Therefore G is n-edge-connected. We show next that G is critically n-edge-connected.

For each vertex u in G, we consider the following two cases for a cutset containing it.

Case 1. u is in a vertex cutset T with $|T| \leqq n - 1$, then $\lambda(G - u) \geqq n - 1$. Since $\lambda(G - u) \leqq \delta(G - u) = n - 1$, we have $\lambda(G - u) = n - 1$.

Case 2. Every vertex cutset containing u has at least n vertices. Suppose $\lambda(G - u) < n - 1$. Let $\hat{E}$ be a minimum edge-cutset of $G - u$, and H_1, H_2 be two components of $(G - u) - \hat{E}$. $|V(H_1)| + |V(H_2)| = m - 1 \geqq (2n + 1) - 1 = 2n$. Without loss of generality, let $|V(H_1)| \geqq n$. Since $|\hat{E}| < n - 1$, u must be adjacent to some vertices in H_1 and some vertices in H_2, as shown in Fig. 5.

Let T_1 be the set of vertices in H_1 which are incident with $\hat{E}$. $|T_1| < n - 1$, since $|\hat{E}| < n - 1$. $|V(H_1) - T_1| > 1$. Thus $T_1 \cup \{u\}$ is a vertex cutset of G and $|T_1 \cup \{u\}| \leqq n - 1$, a contradiction to the assumption of this case. So $\lambda(G - u) \geqq n - 1$. Since $\lambda(G - u) \leqq \delta(G - u) = n - 1$, we have $\lambda(G - u) = n - 1$. Therefore G is critical with respect to $\lambda(G)$.

G is almost regular of degree n, by the proof of Theorem 8, $|E(G)| = \{mn/2\}$, where m is the order of G. Therefore, G is a minimum critically n-edge-connected graph with order m. QED

A vertex u of a graph G is called *critical* if u is contained in a minimum vertex cutset. Thus, we have the following lemma.

LEMMA 12. *A vertex u in graph G is critical if and only if $\kappa(G - u) = \kappa(G) - 1$.*

COROLLARY 13. *For any given positive integers m, n, such that $m \geqq 2n + 1$, $|V(G)| = m$, and $\kappa(G) \geqq n - 1$, $G = (V, E)$ is a minimum critically n-edge-connected graph if and only if G is almost regular of degree n, and for any critical vertex u, $\lambda(G - u) \geqq n - 1$.*

Next, we give some examples to illustrate Theorem 11 and Corollary 13.

Example 1. G is shown in Fig. 6.

G is almost regular of degree 5, $\kappa(G) = 3$. For any vertex u in a vertex cutset T with $|T| \leqq 4$, $\lambda(G - u) \geqq 4$. By Theorem 11, G is a minimum critically 5-edge-connected graph.

Example 2. G is shown in Fig. 7.

G is almost regular of degree 5, $\kappa(G) = 4$, and for any critical vertex u, $\lambda(G - u) \geqq 4$. By Corollary 13, G is a minimum critically 5-edge-connected graph.

Example 3. G is shown in Fig. 8.

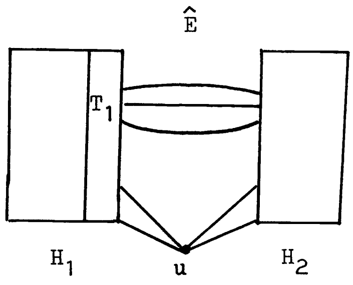


FIG. 5

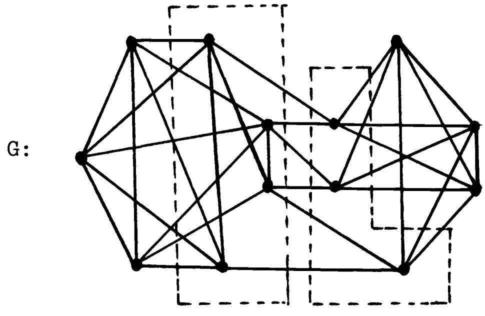


FIG. 6

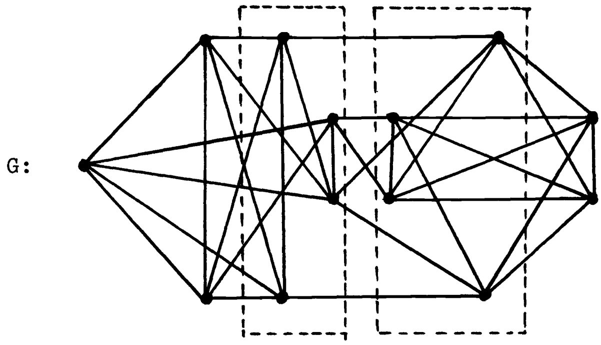


FIG. 7

G is almost regular of degree 5, $\kappa(G) = 3$, u is in a vertex cutset S, $|S| = 4$, $\lambda(G - u) = 3 < 5 - 1$. By Theorem 11 G is not a minimum critically 5-edge-connected graph. In fact, G is not critical with respect to $\lambda(G)$.

Example 4. G is shown in Fig. 9.

G is almost regular of degree 5, $\kappa(G) = 4$, u is a critical vertex, but $\lambda(G - u) = 3 < 5 - 1$. By Corollary 13, G is not a minimum critically 5-edge-connected graph. In fact, G is not critical with respect to $\lambda(G)$.

COROLLARY 14. *For positive integers m, n, $m \geqq 2n + 1$, at least one of n or m is even, and $|V(G)| = m$, $G = (V, E)$ is a minimum critically n-edge-connected graph if and only if G is regular of degree n, and for any vertex u in a vertex cutset T with $|T| \leqq n - 1$, $\lambda(G - u) \geqq n - 1$.*

Proof. By Theorem 11, G is a minimum critically n-edge-connected graph if and only if G is almost regular of degree n, and for any vertex u in a vertex cutset T with $|T| \leqq n - 1$, $\lambda(G - u) \geqq n - 1$.

Now, suppose that G is not regular of degree n, but G is almost regular of degree n. Then $\sum_{v \in V(G)} \deg_G v = n(m - 1) + (n + 1) = nm + 1$ is odd, since nm is even. But $\sum_{v \in V(G)} \deg_G v = 2 \cdot |E|$, so we obtain a contradiction. Conversely, if G is regular of degree n, then G is almost regular of degree n. QED

4. NP-completeness. A problem is in the class NP if some nondeterministic machine could, in every instance, find the answer in a number of steps which is bounded by some fixed polynomial in the length of the input data. A problem is NP-*complete* if it is in NP, and the existence of a deterministic polynomial algorithm, for it would imply the

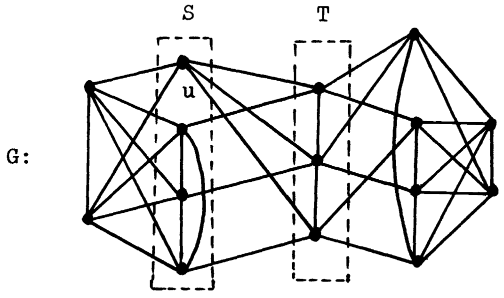


FIG. 8

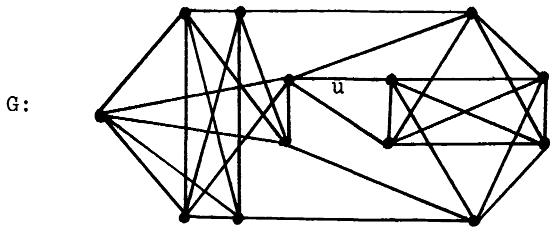


FIG. 9

existence of a deterministic polynomial algorithm for all NP problems. The proof technique for NP-completeness in this section uses the restriction technique. An NP-completeness proof by *restriction* for a given problem $Q \in$ NP consists simply of showing that Q contains a known NP-complete problem R as a special case.

The main problem in this section is as follows:

Problem n-EDGE.

Instance: $G = (V, E)$, a positive integer n, $1 < n \leqq |V| - 1$.

Question: Is there a minimum critically n-edge-connected subgraph $G' = (V, E')$ of G?

We shall show that Problem n-EDGE is NP-complete. To do this, we will use the NP-complete problem, the *Hamiltonian Circuit Problem* (HC).

Problem HC.

Instance: Graph $G = (V, E)$.

Question: Does G contain a Hamiltonian circuit?

LEMMA 15. *$G' = (V, E')$ is a connected spanning subgraph of $G = (V, E)$ and G' is almost regular of degree 2 if and only if G' is a Hamiltonian circuit of G.*

Lemma 15 is proved by using the facts that the number of vertices of odd degree for any graph is even, a connected graph with no vertices of odd degree is Eulerian, and an Eulerian circuit in a 2-regular graph must be a Hamiltonian circuit.

There are many polynomial time algorithms for computing the number of components of a graph $G = (V, E)$ including the one given in [8].

Now we consider Problem ARn.

Problem ARn.

Instance: $G = (V, E)$, a positive integer n, $1 < n \leqq |V| - 1$.

Question: Is there a spanning connected subgraph $G' = (V, E')$, such that G' is almost regular of degree n?

THEOREM 16. *Problem* ARn *is* NP-*complete*.

Proof. First, we prove that Problem ARn is in NP: Given a yes solution (called certificate) to Problem ARn, we give a polynomial checking algorithm:

Certificate: a subgraph G' of G.

CERTIFICATE-CHECKING ALGORITHM (*Procedure* I):

```
 Begin
1. If $V(G') \neq V(G)$
      Then return "No"
      Else
2.        If $c(G')$ (the number of components of $G'$) $\geqq 2$
             Then return "No"
             Else
```

```
3.          Sort degrees of vertices in $G'$, such that
                $d_1 \leqq d_2 \leqq d_3 \leqq \cdots \leqq d_m$;
4.          If $(d_1 = d_2 = d_3 = \cdots = d_{m-1} = n)$ and $(d_m = n$ or $d_m = n + 1)$
               Then return "Yes"
               Else return "No";
 End.
```

Step 2 is a polynomial procedure. Step 3 is a sorting procedure, so it also runs in polynomial time. Therefore, the certificate-checking algorithm runs in polynomial time, Problem ARn is in NP.

Let $n = 2$. Problem ARn is reduced to Problem HC by Lemma 15. So a specified type of instance of Problem ARn is NP-complete. By the "restriction technique," Problem ARn is NP-complete. QED

Problem n-EDGE-*T*.

Instance: $G = (V, E)$, a positive integer n, $1 < |V|/2 \leqq n \leqq |V| - 1$.

Question: Is there a minimum critically n-edge-connected subgraph $G' = (V, E')$ of G?

THEOREM 17. *Problem n*-EDGE-*T is* NP-*complete.*

Proof. By Theorem 8, Problem n-EDGE-T is the same as Problem ARn. So Problem n-EDGE-T is NP-complete. QED

Problem MENS (*Minimum n-edge-connected subgraph*).

Instance: $G = (V, E)$ and positive integers $n \leqq |V|$ and $b \leqq |E|$.

Question: Is there a subset $E' \subseteq E$ with $|E'| \leqq b$ such that $G' = (V, E')$ is n-edge-connected?

COROLLARY 18. *Problem* MENS *is* NP-*complete* [4].

Therefore, if $G' = (V, E')$ is a certificate, then there is a polynomial time certificate-checking algorithm for Problem MENS, we call it "Procedure II."

THEOREM 19. *Problem n*-EDGE *is* NP-*complete.*

Proof. First, we show that Problem n-EDGE is in NP.

Certificate: A subgraph G' of G.

CERTIFICATE-CHECKING ALGORITHM:

```
 Begin
1.  If $G'$ is not a spanning connected subgraph of $G$ or $G'$ is not almost regular of
    degree $n$—(Call Procedure I)
       Then return "No"
       Else
2.        If $G'$ is not $n$-edge-connected—(Call Procedure II)
             Then return "No"
             Else
3.              For I := 1 to $|V|$.
                   Construct $H' = G' - v_I$, $H = G - v_I$;
                      If $H'$ is not $(n - 1)$-edge-connected—(Call Procedure II (Instance:
                      $H, n - 1$))
                         Then return "No" and go to 5.
                         Else go to loop 3;
4.                 Return "Yes";
5.  End.
```

In step 1, Procedure I runs in polynomial time P_1.

In step 2, Procedure II runs in polynomial time P_2.

In step 3, the number of computation steps is $O(P_2 \cdot |V|)$.

Therefore, the certificate-checking algorithm runs in polynomial time, Problem n-EDGE is NP.

If we use instance n, $|V|/2 \leqq n \leqq |V| - 1$, Theorem 17 and the "restriction technique," Problem n-EDGE is NP-complete. QED

We have shown that the problem of finding a minimum critically n-edge-connected spanning subgraph of G is NP-complete. If we place any restrictions on graph G other than the ones imposed in Theorems 8, 10, 11 and Corollary 13 does the problem become easier?

We thank the referee of an earlier version of this paper for his helpful suggestions.

REFERENCES

[1] G. CHARTRAND, *A graph-theoretic approach to a communication problem*, SIAM J. Appl. Math., 14 (1966), pp. 778–781.

[2] M. B. COZZENS AND S.-S. Y. WU, *Maximum critically n-edge-connected graphs*, J. Graph Theory, submitted.

[3] R. C. ENTRINGER, *Characterization of maximum critically* 2*-connected graphs*, J. Graph Theory, 2 (1978), pp. 319–327.

[4] M. R. GAREY AND D. S. JOHNSON, *Computer and Intractability: A Guide to the Theory of* NP*-Completeness*, W. H. Freeman, San Francisco, 1978.

[5] F. HARARY, *The maximum connectivity of a graph*, Proc. Nat. Acad. Sci. U.S.A., 48 (1962), pp. 1142–1146.

[6] ———, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[7] H. J. KROL AND H. J. VELDMAN, *On maximum critically h-connected graphs*, Discrete Math., 52 (1984), pp. 225–234.

[8] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization, Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

013

FOUR VARIATIONAL FORMULATIONS OF THE CONTRAHARMONIC MEAN OF OPERATORS*

W. L. GREEN† AND T. D. MORLEY†

Abstract. The authors establish four variational expressions, each related to the parallel sum, for the contraharmonic mean of two positive operators or matrices.

Key words. contraharmonic mean, parallel sum, positive operator, positive semi-definite matrix, variational formulation

AMS(MOS) subject classifications. 15A45, 49A29, 47D99, 26D20

1. Introduction. The *contraharmonic mean* (see [2]) of two positive scalars is defined by the formula

$$C(a,b)=\frac{a^2+b^2}{a+b}.$$

This may be rewritten as

$$C(a,b)=(a+b)-2(a^{-1}+b^{-1})^{-1},$$

which is twice the arithmetic mean minus the harmonic mean.

In [2] the *contraharmonic mean* of positive semi-definite matrices was defined by

$$C(A,B)=A+B-2(A{:}B), \tag{1}$$

where $A{:}B$ denotes the *parallel sum* of Anderson and Duffin (see [1] and [4]), given by

$$A{:}B=\lim_{\varepsilon\downarrow 0} A(A+B+\varepsilon I)^{-1}B. \tag{2}$$

The formula in (2) above has proved to be a satisfactory definition of the parallel sum for positive operators on a Hilbert space (see for example [4]), and we therefore define the contraharmonic mean for such operators by means of (1) and (2). (The authors of [2] are well aware of the extendibility of some of their results to infinite dimensions.) In this paper we give several variational definitions of the contraharmonic mean of two positive operators.

2. Preliminaries. Let $\mathcal{X}$ be a (complex) Hilbert space, with inner product $(\cdot,\cdot)$. A bounded linear operator A: $\mathcal{X}\to\mathcal{X}$ is termed positive if $(Ax,x)\geqq 0$ for all $x\in\mathcal{X}$. If A and B are operators, then we write $A\geqq B$ to mean A-B is positive. This defines a partial order on the set of bounded linear operators on $\mathcal{X}$. If A and B are positive operators, we define their *parallel sum*, $A{:}B$, by

$$A{:}B=\lim_{\varepsilon\downarrow 0} A(A+B+\varepsilon I)^{-1}B.$$

The parallel sum has been extensively studied by Anderson, Ando, Duffin, Fillmore, Mitra, Puri, Trapp, Williams and many others, [1], [4]–[10]. The following theorem summarizes some of the results from [4].

* Received by the editors December 29, 1986; accepted for publication (in revised form) April 21, 1987.

† School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

THEOREM 1. *Let A and B be positive operators; then*

(a) $A{:}B$ *is positive,*

(b) $(A{:}Bc, c) = \inf_{x+y=c} (Ax, x) + (By, y),$

(c) $$A{:}B = \sup \left\{X\colon \begin{bmatrix} A-X & A \\ A & A+B \end{bmatrix} \geqq 0 \textit{ and } X \geqq 0\right\}.$$

(In part (c) the sup is with respect to the partial order defined above.)

3. Four variational formulations. In this section we give four variational formulas for the contraharmonic mean $C(A, B)$ of two positive operators,

$$\begin{aligned} C(A,B) &= A+B-2(A{:}B) \\ &= \lim_{\varepsilon \downarrow 0} A+B-2A(A+B+\varepsilon I)^{-1}B. \end{aligned}$$

THEOREM 2. *Let A and B be positive operators; then*

$$(C(A,B)c, c) = \sup_{x+y=c} ((A-B)x, x) + ((B-A)y, y) + 2 \operatorname{Re} ((A+B)x, y).$$

Proof. By definition of the contraharmonic mean, we have

$$(C(A,B)c, c) = ((A+B)c, c) - 2((A{:}B)c, c).$$

By Theorem 1(b) this is equal to

$$\begin{aligned} ((A+B)c, c) - 2 \inf_{x+y=c} (Ay, y) + (Bx, x) &= \sup_{x+y=c} ((A+B)c, c) - 2(Ay, y) - 2(Bx, x) \\ &= \sup_{x+y=c} ((A+B)(x+y), x+y) \\ &\quad - 2(Ay, y) - 2(Bx, x). \end{aligned}$$

Expanding the first term and rearranging things a bit gives the result. □

COROLLARY 3. *Let A and B be as above. Then*

$$(C(A,B)c, c) = \sup_{x+y=c} \left(\begin{bmatrix} A-B & A+B \\ A+B & B-A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right).$$

Proof. Expanding the inner product, we obtain

$$\left(\begin{bmatrix} A-B & A+B \\ A+B & B-A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right) = ((A-B)x, x) + ((B-A)y, y) + 2 \operatorname{Re} ((A+B)x, y),$$

and the result now follows from the previous result. □

THEOREM 4. *Let A and B be positive operators. Then*

$$C(A,B) = \inf \left\{ X \geqq 0\colon \begin{bmatrix} X-(A-B) & X-(A+B) \\ X-(A+B) & X-(B-A) \end{bmatrix} \geqq 0 \right\}.$$

Proof. Let X be positive; then

$$\left(\begin{bmatrix} X & X \\ X & X \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right) = (Xx, x) + (Xy, y) + 2 \operatorname{Re} (Xx, y) = (X(x+y), x+y).$$

Thus, if we set $C = C(A, B)$, we have by Corollary 3 that

$$\begin{bmatrix} C & C \\ C & C \end{bmatrix} \geqq \begin{bmatrix} A-B & A+B \\ A+B & B-A \end{bmatrix}.$$

This says that if we set

$$\mathscr{C}(A,B) = \left\{ X: \begin{bmatrix} X-(A-B) & X-(A+B) \\ X-(A+B) & X-(B-A) \end{bmatrix} \geqq 0, X \geqq 0 \right\},$$

then $C(A, B) \in \mathscr{C}(A, B)$.

To complete the proof we must show that if $X \in \mathscr{C}(A, B)$, then $X \geqq C(A, B)$. So let $X \in \mathscr{C}(A, B)$; then as above

$$(X(x+y), x+y) \geqq \left(\begin{bmatrix} A-B & A+B \\ A+B & B-A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right).$$

Letting $c = x + y$, maximizing the right-hand side over all $x + y = c$, and applying Corollary 3, we get

$$(Xc, c) \geqq C(A, B),$$

and the result follows. □

THEOREM 5. *Let A and B be positive; then*

$$C(A,B) = \inf \left\{ X: \begin{bmatrix} A-B+X & 2A \\ 2A & 2(A+B) \end{bmatrix} \geqq 0, X \leqq A+B \right\}.$$

Proof. By Theorem 1(c),

$$2(A:B) = \sup \left\{ X \middle| \begin{bmatrix} 2A-X & 2A \\ 2A & 2(A+B) \end{bmatrix} \geqq 0, X \geqq 0 \right\}.$$

If we set $C = A + B - X$, then

$$C(A,B) = \inf \left\{ C: X \geqq 0, X = A+B-C, \begin{bmatrix} 2A-X & 2A \\ 2A & 2(A+B) \end{bmatrix} \geqq 0 \right\}.$$

The condition $X \geqq 0$ translates into $C \leqq A + B$, and thus

$$C(A,B) = \inf \left\{ C \geqq 0: C \leqq A+B, \begin{bmatrix} 2A-(A+B-C) & 2A \\ 2A & 2(A+B) \end{bmatrix} \right\},$$

and the result follows. □

4. Comments. The authors of the present paper happened upon the contraharmonic mean quite independently of [2]. We were studying (see [7]) the formula from Fillmore and Williams [6] for the parallel sum of two operators:

THEOREM [6]. *Let A and B be positive operators. Then there are unique operators D and E such that*

(a) $A^{1/2} = (A + B)^{1/2}D$,

(b) $B^{1/2} = (A + B)^{1/2}E$,

(c) $\ker D^* \supseteq \ker (A + B)^{1/2}$,

(d) $\ker E^* \supseteq \ker (A + B)^{1/2}$.

Moreover if D and E are as above, then

$$A:B = A^{1/2}D^*EB^{1/2}.$$

Proof. See [6]. □

If D and E are as above, then it can easily be shown that

$$C(A,B)=A^{1/2}D^*DA^{1/2}+B^{1/2}E^*EB^{1/2}.$$

REFERENCES

[1] W. N. Anderson and R. J. Duffin, *Series and parallel additions of matrices*, J. Math. Anal. Appl., 26 (1969), pp. 576–594.

[2] W. N. Anderson, M. E. Mays and G. E. Trapp, *The contraharmonic mean of HSD matrices and electrical networks*, Proc. 27th Midwest Symposium on Circuits and Systems, June 1984, pp. 665–668.

[3] W. N. Anderson, M. E. Mays, T. D. Morley and G. E. Trapp, *The contraharmonic mean of hermitian semidefinite matrices*, this Journal, 8 (1987), pp. 674–682.

[4] W. N. Anderson and G. E. Trapp, *Shorted operators* II, SIAM J. Appl. Math., 28 (1975), pp. 60–71.

[5] T. Ando, *Topics in operator inequalities*, Lecture Notes, Hokkaido University, Sapporo, Japan, 1978.

[6] P. A. Fillmore and J. P. Williams, *On operator ranges*, Adv. in Math., 7 (1971), pp. 254–281.

[7] W. L. Green and T. D. Morley, *Operator means, fixed points and the norm convergence of monotone approximants*, Math. Scand., to appear.

[8] T. D. Morley, *Parallel summation, Maxwell's principles and the infimum of projections*, J. Math. Anal. Appl., 70 (1979), pp. 33–41.

[9] S. K. Mitra and M. L. Puri, *On parallel sum and difference of matrices*, J. Math. Anal. Appl., 44 (1973), pp. 92–97.

[10] G. E. Trapp, *Hermitian semidefinite matrix means and inequalities—an introduction*, Linear and Multilinear Algebra, 16 (1984), pp. 113–123.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

014

THE CONTRAHARMONIC MEAN OF HSD MATRICES*

WILLIAM N. ANDERSON, JR.†, MICHAEL E. MAYS‡, THOMAS D. MORLEY§ AND GEORGE E. TRAPP¶

Abstract. For positive scalars a and b the contraharmonic mean of a and b, $C(a, b)$, is defined by

$$C(a,b)=(a^2+b^2)/(a+b).$$

In this paper we consider a natural matrix generalization of the contraharmonic mean, fit this into the matrix analogue of some of the classical scalar inequalities for means, develop computational procedures which let us generate the matrix analogues of an infinite family of scalar means, and study fixed point problems. Finally, we mention a relationship between least squares problems and the contraharmonic mean.

Key words. harmonic, arithmetic, contraharmonic

AMS(MOS) subject classifications. 15A24, 15A45

1. Introduction.

If A and B are Hermitian positive semidefinite matrices we define

$$C(A,B)=A+B-2(A:B),$$

where $A : B$ is the operation of parallel addition introduced by Anderson and Duffin [1]. Many of the properties of $C(A, B)$ are related to those of the harmonic mean $2(A : B)$ and the arithmetic mean $(A + B)/2$. For example,

$$A+B\geqq C(A,B)\geqq (A+B)/2$$

and

$$(C(A,B)+2(A:B))/2=(A+B)/2.$$

The dual mean $C'(A, B) = C(A^{-1}, B^{-1})^{-1}$ can be written as

$$C'(A,B)=(A:B)+2(A:B)C(A,B)^{-1}(A:B),$$

or as

$$C'(A,B)=[A(A:B)^{-1}A]:[B(A:B)^{-1}B].$$

We will also present variational characterizations of both the contraharmonic mean and its dual. These variational characterizations arise from the representation of the contraharmonic mean using the operation of parallel subtraction.

The equation

$$C(A,X)=A+B$$

is equivalent to the fixed point problem

$$X=2(A:X)+B.$$

If B is positive definite this latter problem has a unique solution according to the work of Anderson, Kleindorfer, Kleindorfer and Woodroofe [2], and in fact is the same as the solution to

$$C(Y,B)=A+B.$$

* Received by the editors December 29, 1986; accepted for publication (in revised form) May 8, 1987.

† Department of Mathematics and Computer Science, Fairleigh Dickinson University, Teaneck, New Jersey 07666.

‡ Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506.

§ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

¶ Department of Statistics and Computer Science, West Virginia University, Morgantown, West Virginia 26506.

In the scalar case the solution would be

$$x=\{(a+b)+(a^2+6ab+b^2)^{1/2}\}/2,$$

but in an HSD setting the fractional power needs to be interpreted using the geometric mean operation #, defined by Pusz and Woronowicz [14] as

$$A\#B=A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}.$$

Another interesting problem is that of inverse means: given HSD matrices E and F and means M_1 and M_2, when can we find HSD matrices A and B satisfying $E = M_1(A, B)$ and $F = M_2(A, B)$? We consider this problem when one of the specified means is the contraharmonic mean.

In the final section we exhibit a relationship between some classical least squares problems and the contraharmonic mean. We show that there are two natural ways to define the contraharmonic mean of three HSD matrices.

2. Preliminaries. The contraharmonic mean is one of the classical means of Greek mathematics. Its name arises from the fact that, just as solving for x in the equation

$$(a-x)/(x-b)=a/b$$

yields the harmonic mean, so solving for x in the equation

$$(a-x)/(x-b)=b/a$$

yields

$$x=C(a,b)=(a^2+b^2)/(a+b).$$

Many investigators have linked means together by introducing one or more parameters into the definition which, when varied, generate some of the "named" means. Thus the harmonic mean is the case $s = 0$ and the contraharmonic mean is the case $s = 2$ for the mean studied by Gini [8], Beckenbach [6] and Lehmer [9]:

$$G_s(a,b)=(a^s+b^s)/(a^{s-1}+b^{s-1}).$$

The arithmetic mean and geometric mean can also be associated with this mean, by letting $s = 1$ and $s = \frac{1}{2}$. Other parametrizations are discussed in Mays [11].

To generalize from the scalar case, we consider matrices on a finite-dimensional inner product space. The inner product is denoted by $\langle\, ,\rangle$. A matrix is called Hermitian positive semidefinite (HSD) if $A = A^*$, where A^* is the adjoint (conjugate transpose) of A, and $\langle Ax, x\rangle \geqq 0$ for all vectors x. If A and B are both HSD, we write $A \geqq B$ if $A - B$ is HSD.

To generalize the contraharmonic mean to the case of HSD matrices, we require three special HSD matrix operations: the parallel sum (harmonic mean), the geometric mean and the parallel difference. If A and B are invertible HSD matrices, the parallel sum, denoted $A : B$, and the geometric mean, denoted $A \# B$, are defined by

$$A:B=A(A+B)^{-1}B$$

and

$$A\#B=A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}.$$

In the case that the inverses do not exist, the limit of nonsingular approximations may be used as the definition.

Another operation that we will use is a subtraction operation related to $A : B$. We seek HSD X so that $A : X = C$ for given A and C. We refer to [3] and [13] for additional background; here we only require the following result from [13].

If there is an X so that $A : X = C$, then one such X is given by the variational formula

$$\min \left\{ X \mid X \text{ is HSD and } \begin{bmatrix} A-C & A \\ A & A+X \end{bmatrix} \text{ is HSD} \right\}.$$

The minimum X is called the parallel difference of C and A and written $C \div A$. The minimum X is the only solution of $A : X = C$ such that the range of X is contained in the range of A. We use the parallel difference to characterize the dual of the contraharmonic mean. This X may also be obtained from the formula

$$X = A(A - C)^{+}C$$

where the superscript "+" denotes the Moore–Penrose generalized inverse.

A fundamental theorem used in our investigation is the arithmetic-geometric-harmonic inequality

$$(A + B)/2 \geqq A \# B \geqq 2(A : B),$$

which is valid for all HSD matrices A and B. Trapp [15] has more information and background on these operations and inequalities.

3. The contraharmonic mean and its dual.

If A and B are HSD matrices, we define the contraharmonic mean, denoted $C(A, B)$, by

$$C(A, B) = A + B - 2(A : B). \tag{1}$$

This definition is motivated by an identity in the scalar case. Since both + and : are commutative, this operation is commutative. When Q^{-1} exists we can write

$$Q(C(A, B))Q^* = C(QAQ^*, QBQ^*),$$

a matrix homogeneity property stronger than the standard requirement for homogeneity, in which Q would have to be a scalar.

Clearly $A + B \geqq C(A, B)$, and since $(A + B)/2 \geqq 2(A : B)$ we see that

$$C(A, B) \geqq (A + B)/2.$$

This guarantees that $C(A, B)$ is HSD.

A direct computation shows that

$$(C(A, B) + 2(A : B))/2 = (A + B)/2,$$

i.e., the arithmetic mean of the contraharmonic and harmonic means is the arithmetic mean. When $A + B$ is invertible, another direct computation using the equivalent parallel formula

$$A : B = A - A(A + B)^{-1}A$$

yields

$$(A + B)/2 - 2(A : B) = (A - B)(A + B)^{-1}(A - B)/2;$$

hence

$$C(A, B) = (A + B)/2 + (A - B)(A + B)^{-1}(A - B)/2.$$

As a corollary, note that

$$(A + B)/2 \geqq (A - B)(A + B)^{-1}(A - B)/2.$$

because

$$A + B = (A + B)/2 + (A + B)/2 \geqq C(A, B).$$

Duality is a natural matrix mean concept. The dual of the contraharmonic mean, which we will denote $C'(A, B)$, is defined by

$$C'(A,B) = C(A^{-1}, B^{-1})^{-1}$$

when A and B are invertible.

THEOREM 1. $C'(A, B) = A : B + 2(A : B)C(A, B)^{-1}(A : B)$.

Proof. Let $D = A^{-1/2}BA^{-1/2}$. Then

$$\begin{aligned} C(A,B) &= C(A^{1/2}IA^{1/2}, A^{1/2}DA^{1/2}) \\ &= A^{1/2}C(I,D)A^{1/2} \end{aligned}$$

and

$$C(A^{-1},B^{-1})^{-1} = A^{1/2}C(I,D^{-1})^{-1}A^{1/2}.$$

Since I and D commute, as in the scalar case we have

$$C(I,D^{-1})^{-1} = I:D + 2(I:D)C(I,D)^{-1}(I:D).$$

Thus

$$\begin{aligned} C(A^{-1},B^{-1})^{-1} &= A^{1/2}\{I:D + 2(I:D)C(I,D)^{-1}(I:D)\}A^{1/2} \\ &= A^{1/2}(I:D)A^{1/2} \\ &\quad +2A^{1/2}(I:D)A^{1/2}A^{-1/2}C(I,D)^{-1}A^{-1/2}A^{1/2}(I:D)A^{1/2} \\ &= A:B + 2(A:B)(A^{1/2}C(I,D)A^{1/2})^{-1}(A:B) \\ &= A:B + 2(A:B)(C(A,B))^{-1}(A:B), \end{aligned}$$

as desired.

Another representation of the dual is given by the following theorem:

THEOREM 2. $C'(A, B) = (A(A : B)^{-1}A) : (B(A : B)^{-1}B)$.

Proof.

$$\begin{aligned} C'(A,B) &= (A^{-1} + B^{-1} - 2(A^{-1}:B^{-1}))^{-1} \\ &= (A^{-1} - A^{-1}:B^{-1} + B^{-1} - A^{-1}:B^{-1})^{-1}. \end{aligned}$$

Now let

$$X^{-1} = A^{-1} - (A^{-1}:B^{-1}) = A^{-1} - (A+B)^{-1}$$

and

$$Y^{-1} = B^{-1} - (A^{-1}:B^{-1}) = B^{-1} - (A+B)^{-1}.$$

The proof is complete upon noting that $C'(A, B) = X : Y$, and that

$$AX^{-1}A = A - A(A+B)^{-1}A = A:B$$

implies $X = A(A : B)^{-1}A$, with a similar result for Y.

Now we begin with the original formula for the contraharmonic mean and take the duals of each side to generate the following sequence of equations. Note we are using the fact that the arithmetic and harmonic means are duals.

$$C(A,B) = A + B - 2(A:B)$$

or

$$2(A:B) + C(A,B) = A + B.$$

Taking duals yields

$$((A+B)/2):C'(A,B) = A:B,$$

and thus $C'(A, B)$ exists for all invertible A and B. Since the range of $(A + B)/2$ contains the range of $A : B$, according to a result of [3], the equation $((A + B)/2) : X = A : B$ has a solution, and hence a minimum solution. This minimum solution should be defined as $C'(A, B)$ in the most general situation; it is in accordance with $C(A^{-1}, B^{-1})$ for invertible A and B. Since the range of the contraharmonic mean is contained in the range of the arithmetic mean, the dual contraharmonic mean may be written

$$C'(A,B) = (A:B) \div (A+B)/2.$$

Alternatively, we can use the variational characterization of parallel subtraction to write

$$C'(A,B) = \inf\left\{X \mid X \text{ is HSD and } \begin{bmatrix} (A+B)/2 - A:B & (A+B)/2 \\ (A+B)/2 & (A+B)/2+X \end{bmatrix} \text{ is HSD}\right\}.$$

We multiply the composite matrix by 2 (which does not change its HSD character) to obtain the following theorem.

THEOREM 3.

$$C'(A,B) = \inf\left\{X \mid X \text{ is HSD and } \begin{bmatrix} C(A,B) & A+B \\ A+B & A+B+2X \end{bmatrix} \text{ is HSD}\right\}.$$

Green and Morley [7] have developed other variational representations of the contraharmonic and the dual contraharmonic means. We list their results here for completeness and comparison. The proofs are in [7].

THEOREM 4. (a)

$$\langle C(A,B)z, z\rangle = \sup_{x+y=z} \{\langle (A-B)x, x\rangle + \langle (B-A)y, y\rangle + 2\langle (A+B)x, y\rangle\}.$$

(b)

$$\langle C(A,B)z, z\rangle = \sup_{x+y=z} \left\{\left\langle \begin{bmatrix} A-B & A+B \\ A+B & B-A \end{bmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix}\right\rangle\right\}.$$

(c)

$$C(A,B) = \inf\left\{X \geqq 0 \;\middle|\; \begin{bmatrix} X-(A-B) & X-(A+B) \\ X-(A+B) & X-(B-A) \end{bmatrix} \geqq 0\right\}.$$

4. Fixed point problems. We now consider two fixed point problems. Given the HSD matrices A and B, find X and Y so that

$$C(A,X) = A+B, \tag{2}$$

$$C(Y,B) = A+B. \tag{3}$$

The two fixed point problems may be rewritten as

$$X = 2(A:X) + B, \tag{2'}$$

$$Y = 2(B:Y) + A. \tag{3'}$$

Equations (2′) and (3′) are special cases of a fixed point problem studied by Anderson, Kleindorfer, Kleindorfer and Woodroofe [2]. Their form was

$$Z = Q(M:Z)Q^* + N, \tag{4}$$

and they showed that (4) has a unique HSD solution when N is invertible. Assuming A and B are invertible, we have that (2′) and (3′) have unique solutions. We now show that the solutions must be equal. Multiply (2′) by $A^{-1/2}$ on each side. Then for

$$X' = A^{-1/2}XA^{-1/2}$$

and

$$B' = A^{-1/2} B A^{-1/2},$$

we have

$$X' = 2(I : X') + B'. \tag{5}$$

Here we use again the fact that

$$Q(A:B)Q^* = QAQ^* : QBQ^*.$$

With the problem reduced to (5), we can use the fact that I and B' commute to treat (5) as a scalar problem and write

$$X' = (I + B')/2 + SQRT\{(B')^2 + 6B' + I\}.$$

Therefore

$$X = A^{1/2} X' A^{1/2} = (A + B)/2 + A^{1/2} SQRT\{(B')^2 + 6B' + I\} A^{1/2}.$$

The term inside $SQRT$ may be factored as

$$(aB' + bI)(B'/a + I/b)$$

for any scalars a and b such that $a/b + b/a = 6$. Since we wish to keep all matrices HSD we will also require that $a > 0$ and $b > 0$. The $SQRT$ term may then be written as

$$(aB' + bI) \# (B'/a + I/b).$$

Then using the fact that

$$Q(A \# B)Q^* = QAQ^* \# QBQ^*$$

we can write

$$A^{1/2}(aB' + bI) \# (B'/a + I/b) A^{1/2} = (aB + bA) \# (B/a + A/b).$$

Therefore

$$X = (A + B)/2 + (aB + bA) \# (B/a + A/b).$$

A similar analysis is possible for (3′), and we see that Y has an analogous form. The only question that remains is whether the expression

$$(aB + bA) \# (B/a + A/b)$$

depends on the choice of a and b. That the expression does not depend on this choice follows immediately from the results in [2], since the solution is unique.

We close this section by noting some properties of the fixed point viewed as a function of A and B. Let

$$T(A, B) = (A + B)/2 + \{(A + bB) \# (A + B/b)\}/2,$$

where

$$b + 1/b = k \geqq 2.$$

Then

$$T(A, B) \geqq A + B$$

because the work of Ando [5] gives that

$$\{(A + bB) \# (A + B/b)\}/2 \geqq (A \# A + bB \# B/b)/2 = (A + B)/2.$$

We also have that

$$T(0, B) = B,$$

and if

$$A_1 \geqq A_2 \quad \text{and} \quad B_1 \geqq B_2,$$

then

$$T(A_1, B_1) \geqq T(A_2, B_2).$$

5. Loewner's theorem and monotonicity. Given a binary operation & on HSD matrices, define a function $F(z)$ by

$$F(z)I = zI \& I.$$

Such a function $F(z)$ is called a "Pick function." Loewner [10] shows that under certain restrictions on & the operation I & B is monotone if and only if $F(z)$ is analytic in the upper half plane and satisfies Im $(F(z)) > 0$ for Im $(z) > 0$.

For example, parallel addition is monotone. In this case

$$F(z)I = zI : I = (z/(1+z))I.$$

We have for $z = a + bi$ that

$$\text{Im}\,(F(z)) = b/((1+a)^2 + b^2),$$

and the result holds.

For the contraharmonic mean $C(A, B)$,

$$F(z) = (1+z^2)/(1+z).$$

Since this function has a zero at $z = i$, it will have a sign change and we can conclude that $C(A, B)$ is not monotone.

6. Related means of HSD matrices. We have seen that several scalar means in the family

$$G_s(a,b) = (a^s + b^s)/(a^{s-1} + b^{s-1})$$

have interpretations as means of HSD matrices. A set of scalar means arising from a graphical representation of Moskovitz [12],

$$M_s(a,b) = (ab^s + ba^s)/(a^s + b^s),$$

may be interpreted in this way as well.

Three algebraic identities connecting these means provide a recurrence relation that easily extends to the HSD matrix case while avoiding problems of commutativity, so that we have a family of HSD matrix means

$$G_s(A,B)$$

defined for s, an arbitrary integer. These identities are

$$G_s(A,B) = M_s(A^{-1}, B^{-1})^{-1}, \tag{6}$$

$$M_s(A,B) + G_{s+1}(A,B) = A + B \tag{7}$$

and

$$M_s(A,B) = G_{1-s}(A,B). \tag{8}$$

Thus we have means and their duals arising not only for $s = 2$ (the contraharmonic mean) but for $s = 3, 4, \cdots$ as well. In none of these cases, however, is monotonicity preserved.

To see why $G_s(A, B)$ is HSD if A and B are HSD, note that for both symmetry and positivity we may use induction. First,

$$A + B \geqq G_2(A,B) \geqq (A+B)/2,$$

and if

$$A + B \geqq G_s(A,B) \geqq (A+B)/2$$

for all HSD A and B then

$$A^{-1} + B^{-1} \geqq G_s(A^{-1}, B^{-1}) \geqq (A^{-1} + B^{-1})/2,$$

so

$$-(A^{-1}+B^{-1})^{-1} \geqq -G_s(A^{-1},B^{-1})^{-1} \geqq -((A^{-1}+B^{-1})/2)^{-1}.$$

Now add $A + B$ to each part, and note that the left-hand part is bounded above by $A + B$, the right-hand part is $C(A, B)$, which is bounded below by $(A + B)/2$, and the middle is $G_{s+1}(A, B)$, by (6) and (7).

7. Inverse mean problems. In this section we are interested in the following type of problem: Given E and F HSD, when can we find HSD A and B so that $E = (A + B)/2$ and $F = 2(A : B)$? This question and similar questions involving the geometric mean are answered in [4]; here we wish to consider questions of the same form involving the contraharmonic mean. For example, let E and F be HSD. When can we find HSD A and B so that $E = A + B$ and $F = C(A, B)$?

We know that such a representation can exist only when E and F satisfy $2F \geqq E \geqq F$. In fact this inequality is also sufficient for such a representation to exist. This result is presented in the next theorem.

THEOREM 5. *Given HSD E and F with $2F \geqq E \geqq F$, let A and B be defined by*

$$A=(E+E\#(2F-E))/2,$$

$$B=(E-E\#(2F-E))/2.$$

Then A and B are HSD, $A + B = E$, and $C(A, B) = F$.

Proof. It is obvious that A is HSD and $A+B=E$. Since $E \geqq F$, we have $E \geqq 2F - E$ and hence

$$E=E\#E \geqq E\#(2F-E)$$

and B is HSD. To complete the proof we need that $C(A, B) = F$.

Let $X = E \# (2F - E)$. Recall from [3] that $2F - E$ is then equal to $XE^{-1}X$. Write

$$\begin{aligned} C(A,B) &= A+B-2(A:B) \\ &= E-2((E+X)/2:(E-X)/2) \\ &= E-(E+X)(2E)^{-1}(E-X) \\ &= E-(E-XE^{-1}X)/2 \\ &= E-(E-(2F-E))/2 \\ &= F. \end{aligned}$$

Similar techniques lead to the following.

THEOREM 6. *Given HSD E and F with $2F \geqq E \geqq F$, the following A and B are HSD with $E = C(A, B)$ and $F = (A + B)/2$:*

$$A=F+F\#(E-F),$$

$$B=F-F\#(E-F).$$

We leave unresolved the question of finding A and B such that $E = C(A, B)$ and $F = A \# B$. Even the scalar version of this problem is difficult to solve because a quartic equation arises.

8. The contraharmonic mean and least squares problems. If we wish to solve the system $\substack{A\underline{x}=\underline{d}\\ B\underline{x}=\underline{d}}$ as a least squares problem, $\underline{x} = (A^2 + B^2)^{-1}(A + B)\underline{d}$. If instead we use $\underline{x} = C(A, B)^{-1}\underline{d}$, these agree when there is an $\underline{x}$ which simultaneously satisfies the equations $A\underline{x} = \underline{d}$ and $B\underline{x} = \underline{d}$. The two forms also agree if $AB = BA$.

A natural extension of the contraharmonic mean to three variables would seem to be via a least squares problem for a system with three components:

$$A_1\underline{x}=\underline{d},$$

$$A_2\underline{x}=\underline{d},$$

$$A_3\underline{x}=\underline{d}.$$

The least squares solution of this system is

$$\underline{x}=(A_1^2+A_2^2+A_3^2)^{-1}(A_1+A_2+A_3)\underline{d}.$$

In the scalar case,

$$\frac{a_1^2+a_2^2+a_3^2}{a_1+a_2+a_3}=a_1+a_2+a_3-2\frac{a_1a_2+a_1a_3+a_2a_3}{a_1+a_2+a_3}$$

$$=a_1+a_2+a_3-2M.$$

There is more than one analogous identity in the matrix case because there is more than one matrix generalization of M. It is noted in [15] that the following are not equivalent in general:

$$M_1=((A_1:(A_2+A_3))+(A_2:(A_1+A_3))+(A_3:(A_1+A_2)))/2$$

and

$$M_2=2((A_1+(A_2:A_3)):(A_2+(A_1:A_3)):(A_3+(A_1:A_2))).$$

Therefore there are at least these two candidates for the contraharmonic mean of three HSD matrices:

$$C_1(A_1,A_2,A_3)=A_1+A_2+A_3-2M_1$$

and

$$C_2(A_1,A_2,A_3)=A_1+A_2+A_3-2M_2.$$

REFERENCES

[1] W. N. ANDERSON, JR. AND R. J. DUFFIN, *Series and parallel addition of matrices*, J. Math. Anal. Appl., 26 (1969), pp. 576–594.

[2] W. N. ANDERSON, JR., G. D. KLEINDORFER, P. R. KLEINDORFER AND M. B. WOODROOFE, *Consistent estimates of the parameters of a linear system*, Ann. Math. Statist., 40 (1969), pp. 2064–2075.

[3] W. N. ANDERSON, JR., T. D. MORLEY AND G. E. TRAPP, *A characterization of parallel subtraction*, Proc. Natl. Acad. Sci. U.S.A., 76 (1979), pp. 3599-3601.

[4] W. N. ANDERSON, JR. AND G. E. TRAPP, *Inverse problems for means of matrices*, this Journal, 7 (1986), pp. 188-192.

[5] T. ANDO, *Topics on operator inequalities*, Lecture notes, Sapporo, Japan, 1978.

[6] E. F. BECKENBACH, *A class of mean value functions*, Amer. Math. Monthly, 57 (1950), pp. 1–6.

[7] W. L. GREEN AND T. D. MORLEY, *Four variational formulations of the contraharmonic mean of operators*, this Journal, 8 (1987), pp. 670–673.

[8] C. GINI, *Di una formula comprensiva delle medie*, Metron, 13 (1938), pp. 3–22.

[9] D. H. LEHMER, *On the compounding of certain means*, J. Math. Anal. Appl., 36 (1971), pp. 183–200.

[10] K. LOEWNER, *Uber monotone Matrixfunktionen*, Math. Z., 38 (1933), pp. 177–216.

[11] M. E. MAYS, *Functions which parametrize means*, Amer. Math. Monthly, 90 (1983), pp. 677–683.

[12] D. MOSKOVITZ, *An alignment chart for various means*, Amer. Math. Monthly, 40 (1933), pp. 592–596.

[13] E. L. PEKAREV AND JU. L. SMULJAN, *Parallel addition and subtraction of operators*, Math. USSR Izvestija, 10 (1976), pp. 351-370.

[14] W. PUSZ AND S. L. WORONOWICZ, *Functional calculus for sesquilinear forms and the purification map*, Rep. Math. Phys., 8 (1975), pp. 159–170.

[15] G. E. TRAPP, *Hermitian semidefinite matrix means and inequalities—an introduction*, Linear and Multilinear Algebra, 16 (1984), pp. 113–123.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

FAST PARALLEL COMPUTATION OF HERMITE AND SMITH FORMS OF POLYNOMIAL MATRICES*

ERICH KALTOFEN†, M. S. KRISHNAMOORTHY† AND B. DAVID SAUNDERS‡

Abstract. Boolean circuits of polynomial size and polylogarithmic depth are given for computing the Hermite and Smith normal forms of polynomial matrices over finite fields and the field of rational numbers. The circuits for the Smith normal form computation are probabilistic ones and also determine very efficient sequential algorithms. Furthermore, we give a polynomial-time deterministic sequential algorithm for the Smith normal form over the rationals. The Smith normal form algorithms are applied to the rational canonical form of matrices over finite fields and the field of rational numbers.

Key words. parallel algorithm, Hermite normal form, Smith normal form, polynomial-time complexity, probabilistic algorithm, matrix normal form, polynomial matrix, invariant factor

AMS(MOS) subject classifications. 15A21, 68Q40, 15A54

1. Introduction.

The main results of this paper establish fast parallel algorithms for computing the Hermite and Smith normal form of matrices with polynomial entries. The Hermite or Smith normal form of a square matrix is generally defined for the case of entries from a principal ideal domain. For example, the entry domain may be the integers or univariate polynomials over a field. The forms are, roughly speaking, a triangularization, respectively a diagonalization, of the input matrix and they are computed entirely within the domain of the entries. Sequential algorithms for computing the forms are known at least since Hermite [7] and Smith [20], but it requires some effort to show that the forms can be computed in polynomial time. We refer to Kannan and Bachem [13] for integer entries and Kannan [12] for polynomial entries. Applications of both forms include solving linear systems over the domain of entries, computing the geometric multiplicities of the eigenvalues of a matrix, computing the invariant factors of a matrix over a field, and others. For discussion of applications see [1] and [18].

We will show that computing the Hermite normal form over $F[x]$, F a field, is $\mathbf{NC}^1$ reducible to solving singular linear systems. We refer to Cook [4] for the definitions of the complexity classes **NC** and **RNC** and $\mathbf{NC}^1$ reductions. Since the class **NC** requires us to perform field operations on Boolean circuits, the previous claim is precise only for concrete fields such as **Q** or GF(p), the field with p elements. As a corollary we get from the parallel complexity of linear systems [2] that *HERMITE FORM* over $\mathbf{Q}[x]$ is in $\mathbf{NC}^2$ and *HERMITE FORM* over GF(p)[x] is in $\mathbf{RNC}^2$, where *HERMITE FORM* over D is the problem of computing Hermite normal forms over D. Our parallel reduction is completely different from any of the sequential solutions, discussed, for example, in [13]. Of course, we have Kannan's result that *HERMITE FORM* over $\mathbf{Q}[x]$ is in **P** as a consequence, where **P** is the class of sequential polynomial-time problems.

Second, we will present a probabilistic parallel algorithm for computing the Smith normal form over $F[x]$, that is we establish that *SMITH FORM* over $F[x]$ is in $\mathbf{RNC}^2$. The nature of our probabilistic algorithm is such that with controllably small probability an incorrect result might be returned, as with the fast probabilistic parallel rank algorithm [2]. Since Kannan [12] does not prove that his sequential algorithm for *SMITH FORM*

* Received by the editors April 29, 1985; accepted for publication (in revised form) April 23, 1987. This work was supported by National Science Foundation grant MCS 83-14600.

† Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12181.

‡ Department of Computer and Information Sciences, University of Delaware, Newark, Delaware 19716.

over **Q**$[x]$ runs in polynomial time we will also present another sequential algorithm with which we can establish that *SMITH FORM* over **Q**$[x]$ is in **P**. Neither our probabilistic parallel algorithm nor our deterministic sequential algorithm for the Smith normal form is based on repeated computations of Hermite normal forms as is Kannan and Bachem's algorithm. Our key idea in the parallel algorithm is that though each entry in the Smith normal form is a quotient of two GCDs of possibly exponentially many minors we can quickly produce random linear combinations of these minors whose GCD is with high probability equal to the needed GCD. Unlike our parallel Hermite normal form algorithm our parallel solution for the Smith normal form also provides a practical algorithm superior to previously known methods.

We wish to add two remarks. It is possible to use *HERMITE FORM* over **Q**$[x]$ as a tool for solving linear systems over **Q**$[x]$ in polynomial time. Also, however, the fact that solving linear systems over $F[x_1, \cdots, x_v]$, v fixed, is $\mathbf{NC}^1$ reducible to singular linear systems over F is a consequence of Hermann's [8] degree estimates of Hilbert's [9] reduction. See also the appendix of Mayr and Meyer [15] for several corrections to Hermann's proof. Second, we cannot hope to provide fast parallel algorithms for *HERMITE FORM* over **Z** and *SMITH FORM* over **Z** unless progress is made on computing GCDs of integers in parallel, a problem easily shown to be $\mathbf{NC}^1$ reducible to 2 by 2 Hermite or Smith normal forms over **Z**.

In this paper we will restrict ourselves to nonsingular square input matrices, but we note that there are no great difficulties in generalizing our approach to rectangular inputs of nonmaximal rank.

2. Parallel Hermite normal form computation. In this section we construct an $\mathbf{NC}^1$-reduction from *HERMITE FORM* over $F[x]$, F a field, to singular linear systems over F. But first we present the necessary definitions and lemmas.

A nonsingular n by n matrix H over $F[x]$ is in *Hermite normal form* if it is lower triangular, the diagonal entries are monic and the entries before the diagonal entry in each row are of lower degree than the diagonal entry. It is well known that for every nonsingular matrix A there exists a unique unimodular matrix U and matrix H in Hermite normal form such that $AU = H$. H is referred to as the Hermite normal form of A. It is fairly clear that Hermite [7] knew the uniqueness though he did not offer a proof. In any case, we need the uniqueness in a stronger form than is usually presented, which we will include as Lemma 2.1.

For a matrix A over $F[x]$ let $a_{i,j,k}$ denote the coefficient of x^k in the i, jth entry.

LEMMA 2.1. *Given the n by n nonsingular matrix A over $F[x]$ with entry degrees less than d, and the vector $(d_1, \cdots, d_n)$ of nonnegative integers, consider the system $AP = G$, where G is lower triangular, and more specifically,*

$p_{i,j}$ are polynomials of degree less than $nd + \max_{1 \leq i \leq n} d_i$, whose coefficients are unknowns;

$g_{i,i}$ are monic of degree d_i with lower order coefficients unknowns, and for $i > j$, $g_{i,j}$ are polynomials of degree less than d_i with unknowns as coefficients.

This is a system of linear equations over F in the unknown $p_{i,j,k}$ and $g_{i,j,k}$ for which the following statements hold.

(1) *The system has at least one solution, if and only if each d_i is no less than the degree of the ith diagonal entry of a Hermite normal form of A.*

(2) *If each d_i is exactly the degree of the ith diagonal entry of a Hermite normal form of A, then the system has a unique solution, hence G is the unique Hermite normal form of A and P is unimodular.*

Proof. Let H be a Hermite Normal Form of A and U a unimodular matrix such that $AU = H$.

Suppose G and P solve the system for a given degree vector $(d_1, \cdots, d_n)$. Since U is invertible in $F[x]$, we have $G = AP = HU^{-1}P$. Because G and H are triangular and nonsingular, $U^{-1}P$ must be also. It follows that the degrees d_i must be no less than the degrees of $h_{i,i}$, which proves (1) in one direction.

On the other hand, if for each i, we have $d_i \geqq \deg(h_{i,i})$, let

$$D = \operatorname{diag}(x^{d_1 - \deg(h_{1,1})}, \cdots, x^{d_n - \deg(h_{n,n})}).$$

Then the system is solved with $P = UD$ and $G = HD$. Thus (1) is proved if we can show that this solution is expressible within the degree bound given for P. Since $\det(A)P = \operatorname{adj}(A)G$, the degrees in P are bounded by the degrees in $\operatorname{adj}(A)G$, which are bounded by $(n-1)d + \max_{1 \leqq i \leqq n} d_i$.

It remains to show that the solution is unique (i.e., $G = H$, $P = U$) when $d_i = \deg(h_{i,i})$. Let R denote the lower triangular matrix, $U^{-1}P$. It suffices now to show that if G and H are in Hermite normal form and R is a unimodular lower triangular matrix such that $G = HR$, then $R = I$ (and $G = H$). This we do by induction on n, the size of the matrices. Partition this system so that the upper left block is 1 by 1:

$$\begin{bmatrix} g & 0 \\ g^c & G' \end{bmatrix} = \begin{bmatrix} h & 0 \\ h^c & H' \end{bmatrix} \begin{bmatrix} r & 0 \\ r^c & R' \end{bmatrix}.$$

We see that $g = hr$, $g^c = h^c r + H'r^c$, and $G' = H'R'$. Now G' and H' are in Hermite normal form and R' is unimodular, so by induction, R' is the $n-1$ by $n-1$ identity matrix and $G' = H'$. Also, since g and h are of the same degree and monic, we have $r = 1$ and $g = h$. If any entry in the column vector r^c is nonzero, let i be the index of the first nonzero entry. Then

$$(\dagger) \qquad g_i^c = h_i^c + h'_{i,i} r_i^c.$$

Since $\deg(h_i^c) < \deg(h'_{i,i}) = d_i$, the degree of the right-hand side of (†) is no less than d_i. On the other hand, since $\deg(g_i^c) < \deg(g'_{i,i}) = d_i$, the degree of the left-hand side is strictly less, a contradiction. Hence all entries of r^c are 0, and $g^c = h^c$, which completes the proof. □

We now define the size of a matrix A over $F[x]$. Let A be an n by n matrix of d degree polynomials with coefficients in F representable in l bits. Then size $(A) = n^2dl$, which is the number of bits required to write down A in binary.

LEMMA 2.2. *For $d_i \leqq nd$ the linear system of Lemma 2.1 consists of $O(n^3d)$ equations in $O(n^3d)$ unknowns. Its entries are of size l (0's, 1's, and coefficients of A).* □

Now let *LINEAR SYSTEMS* over F be the problem of computing one solution to the (possibly) singular linear system $Ax = b$, or of indicating that a solution does not exist, given an n by n matrix A and length n column vector of l bit entries from F. Following Cook [4], we say that problem X is NC^1 reducible to problem Y, if there is a uniform family of Boolean circuits for solving X which use oracle circuits to solve Y. For the purpose of defining the depth of such circuits an oracle contributes a depth of $\log(r)$, where r is the fan-in to the oracle. The main theorem of this section now follows.

THEOREM 2.1. *HERMITE FORM over $F[x]$ is NC^1 reducible to LINEAR SYSTEMS over F.*

Proof. We construct our circuit as follows from processing units at three levels.

(1) Let $e = nd \geqq \deg(\det(A))$. The input matrix A is passed to each of $n(e+1)$ processors which work in parallel. They are numbered by pairs (i, j), where $1 \leqq i \leqq n$ and $0 \leqq j \leqq e$. The (i, j) processor constructs from A the appropriate input for a *LINEAR*

SYSTEM circuit over F. This determines if the system as described in Lemma 2.1 can be solved when the degree vector is given by $d_i = j$ and $d_k = e$, for $k \neq i$. If the oracle produces a solution, then *true* is passed to the next step. If the oracle indicates no solution exists, then *false* is passed on. By Lemma 2.1 the (i, j) circuit answers *true* just in case the ith diagonal entry of the Hermite normal form has degree less than or equal to j. The depth of the circuit at this point is $O(\log(\text{size}(A)))$, by Lemma 2.2.

(2) The n circuits numbered 1 through n work in parallel. The ith processor gets input from the $e + 1$ circuits of step 1 numbered $(i, 0)$ to (i, e). Its output, d_i, is the minimum j such that the output of processor (i, j) is *true*. Clearly, these circuits have $O(\log(\text{size}(A)))$ depth and polynomial size.

(3) One processor receives the d_is which are the exact degrees of the diagonal entries of the Hermite normal form. It feeds a *LINEAR SYSTEMS* oracle the system described in Lemma 2.1, and, by part 3, obtains the desired Hermite normal form.

COROLLARY. *HERMITE FORM over* $\mathbf{Q}[x]$ *and over* $\mathrm{GF}(p)[x]$ *is in* $\mathbf{NC}^2$.

Proof. The corollary follows from the fact that *LINEAR SYSTEMS* over $\mathbf{Q}$ or $\mathrm{GF}(p)$ is in $\mathbf{NC}^2$ [2], [3], [10], [16]. □

3. Parallel probabilistic Smith normal form computation. A polynomial matrix S is in *Smith normal form* if it is diagonal, each diagonal entry is monic, and each diagonal entry except the last is a divisor of the succeeding one. If S is equivalent to A, i.e., $A = PSQ$, where P and Q are unimodular, then S is called the Smith normal form of A.

LEMMA 3.1. *Let A be an n by n nonsingular matrix over $F[x]$.*

(1) *There is an n by n matrix S in Smith normal form and unimodular matrices P and Q such that $A = PSQ$.*

(2) *Let s_i^* denote the greatest common divisor of all i by i minors of A. Then the diagonal entries in the Smith normal form of A are $s_{1,1} = s_1^*$, and $s_{i,i} = s_i^*/s_{i-1}^*$, for $i > 1$.*

(3) *Two n by n matrices A and B have the same Smith normal form if and only if they are equivalent.* □

For a proof see Gohberg, Lancaster and Rodman [5] or Newman [17].

Let C_i^n denote all i element subsets of $\{1, \cdots, n\}$ and let $A_{I,J}$, for $I, J \in C_i^n$, denote the minor of A restricted to the rows in I and columns in J. By the above theorem we could compute the Smith normal form of A by computing $s_i^* = \mathrm{GCD}_{I,J \in C_i^n} A_{I,J}$. The problem is that there are exponentially many i by i minors. To overcome this problem we compute two random linear combinations of $A_{I,J}$ whose GCD is likely to be the wanted GCD. These are the principal i by i minors of two randomly selected matrices equivalent to A. The following lemma shows this suffices. Let $1 \cdots i$ denote the set $\{1, \cdots, i\}$.

LEMMA 3.2. *Let A be an n by n matrix over $F[x]$, and let s_i^* be as in* (2) *of Lemma* 3.1. *Let* $\mathbf{F}$ *be the extension of $F[x]$ by $4n^2$ indeterminants,* $\mathbf{F} = F[x][\kappa_{j,k}, \lambda_{j,k}, \mu_{j,k}, \nu_{j,k}]$. *Then there exists a polynomial $\pi_i \in \mathbf{F}$ of total degree no more than $4i^2d$ with the following property. For any n by n matrices R, T, U, V over F, $\pi_i(r_{j,k}, t_{j,k}, u_{j,k}, v_{j,k}) \neq 0$ implies that* $\mathrm{GCD}(B_{1 \cdots i, 1 \cdots i}, C_{1 \cdots i, 1 \cdots i}) = s_i^*$, *where $B = RAT$, $C = UAV$.*

Proof. First let the matrices have indeterminate entries, $\mathbf{R} = (\kappa_{j,k})$, $\mathbf{T} = (\lambda_{j,k})$, $\mathbf{U} = (\mu_{j,k})$ and $\mathbf{V} = (\nu_{j,k})$. In this case, we first show $\mathbf{G} = \mathrm{GCD}(\mathbf{B}_{1 \cdots i, 1 \cdots i}, \mathbf{C}_{1 \cdots i, 1 \cdots i}) = s_i^*$ in $\mathbf{F}[x]$ where $\mathbf{B} = \mathbf{R}A\mathbf{T}$ and $\mathbf{C} = \mathbf{U}A\mathbf{V}$, and $\mathbf{F}$ is F with the indeterminates in $\mathbf{R}$, $\mathbf{T}$, $\mathbf{U}$ and $\mathbf{V}$ adjoined. We observe that s_i^* is the only factor of $\mathbf{B}_{1 \cdots i, 1 \cdots i}$ or $\mathbf{C}_{1 \cdots i, 1 \cdots i}$ which lies in $F[x]$. By the Binet–Cauchy formula,

$$\mathbf{B}_{1 \cdots i, 1 \cdots i} = \sum_{K, L \in C_i^n} \mathbf{R}_{1 \cdots i, K} A_{K,L} \mathbf{T}_{L, 1 \cdots i}$$

and

$$\mathbf{C}_{1\cdots i,1\cdots i} = \sum_{K,L\in C_i^n} \mathbf{U}_{1\cdots i,K} A_{K,L} \mathbf{V}_{L,1\cdots i}.$$

Now, clearly the factor of $\mathbf{B}_{1\cdots i,1\cdots i}$ (or $\mathbf{C}_{1\cdots i,1\cdots i}$) in $F[x]$ must divide each $A_{K,L}$. On the other hand, $\mathbf{B}_{1\cdots i,1\cdots i}$ and $\mathbf{C}_{1\cdots i,1\cdots i}$ have no factor in common in $\mathbf{F}[x]\backslash F[x]$ since each involves a different set of indeterminates. This shows our claim on **G**. We now consider

$$\mathbf{B}^* = \frac{\mathbf{B}_{1\cdots i,1\cdots i}}{s_i^*} \quad \text{and} \quad \mathbf{C}^* = \frac{\mathbf{C}_{1\cdots i,1\cdots i}}{s_i^*}.$$

$\mathbf{B}^*$ and $\mathbf{C}^*$ are relatively prime in $\mathbf{F}[x]$, thus $\pi_i = \text{resultant}_x(\mathbf{B}^*, \mathbf{C}^*)$ is nonzero. If

$$\pi_i(r_{j,k}, t_{j,k}, u_{j,k}, v_{j,k}) \neq 0$$

then the polynomials $\mathbf{B}^*(r_{j,k}, t_{j,k}, u_{j,k}, v_{j,k})$ and $\mathbf{C}^*(r_{j,k}, t_{j,k}, u_{j,k}, v_{j,k})$ in $F[x]$ remain relatively prime. (For the theory of resultants, consult, for example , [21, § 5.8].) Therefore GCD $(B_{1\cdots i,1\cdots i}, C_{1\cdots i,1\cdots i}) = s_i^*$.

It remains to estimate the degree of π_i. Clearly, $\deg_x(\mathbf{B}^*)$, $\deg_x(\mathbf{C}^*) \leqq id$. Their degrees in the other indeterminants are bounded by $2i$; thus the degree of $\pi_i \leqq id \times 2i + id \times 2i = 4i^2d$. □

LEMMA 3.3. *With the notation of the previous lemma, if we select the entries in R, T, U, V randomly from a set $S \subset F$ then the probability*

$$\text{Prob}\,(s_i^* = \text{GCD}\,(B_{1\cdots i,1\cdots i}, C_{1\cdots i,1\cdots i}), \textit{for all } i, 1 \leqq i \leqq n) \geqq 1 - \frac{4n^3d}{\text{cardinality }(S)}.$$

Proof. Let $\pi = \prod_{i=1}^n \pi_i$. We are unlucky only if the randomly selected $r_{j,k}$, $t_{j,k}$, $u_{j,k}$ and $v_{j,k}$ are a zero of π. By a result of Schwartz [19] this happens with probability no more than deg (π)/cardinality (S). The degree estimate for π_i now immediately implies that deg $(\pi) \leqq 4n^3d$. □

We now can prove the following theorem.

THEOREM 3.1. *There is a uniform family of probabilistic circuits of depth $O(\log^2(\text{size}\,(A)/\varepsilon))$ and polynomial size which compute the Smith normal form over $F[x]$ correctly with probability $1 - \varepsilon$. These circuits make $O(n^2 \log(nd/\varepsilon))$ random bit choices. In short, SMITH FORM over* $\mathbf{Q}[x]$ *or* $\text{GF}(p)[x]$ *is in* $\mathbf{RNC}^2$.

Proof. By Lemma 3.3 the problem reduces to matrix multiplications, determinant and GCD computations. These are in $\mathbf{NC}^2$ [2]. We must make our $4n^2$ random choices from a subset S of $\mathbf{Q}$ for which $4n^3d$/cardinality $(S) < \varepsilon$. The integers less in absolute value than $4n^3d/\varepsilon$ will do. These are $O(\log(nd/\varepsilon))$ bit numbers.

If the field is too small to allow choice of a sufficiently large set S, S may be chosen from an extension field. Like GCDs, the Smith normal form is an entirely rational form and thus is unchanged if we compute over an extension of the given field. □

Lemma 3.2 remains true if we replace U by an upper triangular and V by a lower triangular matrix, as well as if we do not randomize B. This saves in both matrix multiplications and number of random bits required.

4. Sequential deterministic Smith normal form computation. The purpose of this section is to establish that *SMITH FORM* over $\mathbf{Q}[x]$ is in **P**. First we note that it is a consequence of Kannan [12] that *SMITH FORM* over $GF(p)[x]$ is in **P**, a result on which we will have to depend. We can assume without loss of generality that our input matrix A has integer coefficients. The following lemma is the key to our argument.

LEMMA 4.1. *Let A be a nonsingular n by n matrix over* $\mathbf{Q}[x]$ *with integer coefficients,* $d = \max\{\deg(a_{i,j}) \mid 1 \leqq i, j \leqq n\}$, $L = \max\{|a_{i,j,k}| \mid 1 \leqq i, j \leqq n, 0 \leqq k \leqq \deg(a_{i,j})\}$, l_A *be the leading coefficient of* $\det(A)$, *and let S be the Smith normal form of A,* $d_i = \deg(s_{i,i})$. *Then for any prime p which does not divide l_A, exactly one of the following two conditions can occur for $\bar{S}$, the Smith normal form of A* mod p.

(1) $S \bmod p = \bar{S}$, *or*

(2) $(d_1, \cdots, d_n) \neq (\bar{d}_1, \cdots, \bar{d}_n)$ *with* $\bar{d}_i = \deg(\bar{s}_{i,i})$.

Furthermore, there exists an integer $B_A \leqq (n(d+1)L)^{3n^3d}$ *such that if p does not divide B_A condition* (1) *must occur.*

Proof. Let $\bar{s}_i^* = \mathrm{GCD}_{J,K \in C_i^n}(A_{J,K} \bmod p)$, $1 \leqq i \leqq n$, $\bar{s}_0^* = 1$. Then by Lemma 3.1, $\bar{s}_{i,i} = \bar{s}_i^*/\bar{s}_{i-1}^*$ for $1 \leqq i \leqq n$. It is clear that $s_i^* \bmod p$ divides $\bar{s}_i^*$. Let $\bar{e}_i = \deg(\bar{s}_i^*)$, $e_i = \deg(s_i^*)$. Then $\bar{e}_i \geqq e_i$, $\bar{e}_0 = e_0 = 0$, $\bar{d}_i = \bar{e}_i - \bar{e}_{i-1}$, $d_i = e_i - e_{i-1}$. Either $\bar{e}_i = e_i$ for all $1 \leqq i \leqq n$ or there is a first i such that $\bar{e}_i > e_i$. In the first case, since s_i^* and $\bar{s}_i^*$ are both monic, we have that $s_i^* \bmod p = \bar{s}_i^*$ and hence $s_{i,i} \bmod p = \bar{s}_{i,i}$. In the later case, we have $\bar{d}_i > d_i$.

It remains to establish a condition under which

$$(\dagger) \qquad (\mathrm{GCD}_{J,K \in C_i^n}(A_{J,K})) \bmod p = \mathrm{GCD}_{J,K \in C_i^n}(A_{J,K} \bmod p)$$

for all $1 \leqq i \leqq n$. First we note that for $A_{J,K} = \sum b_j x^j$, $b_j \in \mathbf{Z}$ and $|b_j| \leqq B = (\sqrt{n}(d+1)L)^n$ (cf. [6, Problem 73-17]). Second, we appeal to the following (cf. [11, Lemma 4]).

PROPOSITION. *If* $f_1, \cdots, f_t \in \mathbf{Q}[x]$ *are polynomials with integer coefficients and* $\deg(f_j) \leqq e$, *then there exists an $\bar{e}$ by $\bar{e}$ determinant* $\Delta \in \mathbf{Z}\backslash\{0\}$, $\bar{e} \leqq 2e$, *whose entries are coefficients of the f_j such that, for any prime p which does not divide* Δ,

$$\mathrm{GCD}_{1 \leqq j \leqq t}(f_j \bmod p) = (\mathrm{GCD}_{1 \leqq j \leqq t}(f_j)) \bmod p.$$

Proof. Let $d(x) = \mathrm{GCD}(f_j)$. For any prime p, it is clear that $d \bmod p$ divides $\mathrm{GCD}(f_j \bmod p)$, since $d \bmod p$ divides each $f_j \bmod p$. We show that the converse holds for most primes. There exist $s_1, \cdots, s_t \in \mathbf{Q}[x]$ with $\deg(s_j) < e$ such that $\mathrm{GCD}(f_j) = \sum f_j s_j$. Since each term has degree at most $e + (e-1)$, this equation may be viewed as a linear system, $d = Fs$ of at most $2e$ equations over $\mathbf{Q}$ in te variables, the coefficients of the s_j. The entries of the matrix F are the coefficients of the f_j. Such a linear system has a solution just in case the rank of F is the same as the rank of the augmented matrix (F, d). Since the system has a solution over $\mathbf{Q}$, the rank condition holds. If the rank of $F \bmod p$ is $\bar{e} \leqq 2e$, then an $\bar{e}$ by $\bar{e}$ minor, Δ, of F must be nonzero. If Δ is nonzero mod p as well, it follows that the rank condition will hold mod p and hence the system will have a solution, s'. Thus $\mathrm{GCD}(f_j \bmod p)$ divides $\sum(f_j \bmod p)(s_j') = d \bmod p$, for polynomials s_j' appropriately constructed from s'. □

Continuing the proof of Lemma 4.1, we apply this proposition to $A_{J,K}$ and obtain as the asserted determinant an integer B_i,

$$B_i \leqq \sqrt{2di}\,B^{2di} \leqq \sqrt{2dn}\,(n(d+1)^2L^2)^{dn^2} < (n(d+1)L)^{3dn^2},$$

such that if p does not divide B_i, $(\dagger)$ is satisfied for i. It remains to set $B_A = \prod_{i=1}^n B_i$. □

The deterministic algorithm is now easy to describe. First we select

$$k = 2\lceil \log_2(l_A(n(d+1)L)^{3n^3d}) \rceil \geqq 2\lceil \log_2(l_A B_A) \rceil$$

primes p_j and compute for all primes not dividing l_A the Smith normal form $\bar{S}_j$ of $A \bmod p_j$. We note that the kth prime p_k is $\leqq k \log(k)$, $k \geqq 6$, which makes this step a polynomial-time process. Also, more than half of the primes considered do not divide $l_A B_A$. Hence by the above lemma a majority of the $\bar{S}_j$ must possess the same diagonal-degree vector, say these mod p_j, $j \in J$. Also by the lemma $\bar{S}_j$, $j \in J$, is an image of S. By Chinese remaindering we compute

$$\tilde{S} \equiv S \bmod \tilde{p}, \qquad \tilde{p} = \prod_{j \in J} p_j.$$

It remains to recover the coefficients $s_{i,i,k}$ from their modular images $\tilde{s}_{i,i,k}$. We first observe that the $s_{i,i}$ are monic factors of det (A) over $\mathbf{Q}[x]$. Therefore by Gauss's lemma the denominators of $s_{i,i,k}$ are factors of l_A and hence relatively prime to $\tilde{p}$. We now claim that

$$s_{i,i,k} = \frac{l_A \tilde{s}_{i,i,j} \bmod \tilde{p}}{l_A},$$

where the modulus in the numerator is taken balanced. The only problem could be that $\tilde{p}$ were too small to capture l_A × the numerator of $s_{i,i,k}$. But the integral coefficients of factors of det (A) are absolutely bounded by $2^{nd}B$ (see [14, § 4.6.2, Exercise 20]). Now clearly $2l_A 2^{nd}B < \tilde{p}$ and we have the following theorem.

THEOREM 4.1. *SMITH FORM over* $\mathbf{Q}[x]$ *is in* **P**.

5. Rational canonical form and similarity.

If A is a matrix over a field F, then the diagonal entries of the Smith normal form of $xI - A$ (over $F[x]$) are the invariant factors of A. The invariant factors characterize A up to similarity and their companion matrices form the diagonal blocks of the rational canonical form R of A. Thus we can compute *RATIONAL FORM* in $\mathbf{RNC}^2$ and in **P**. Furthermore, we can compute the similarity transform U such that $UAU^{-1} = R$, whereas for the Smith normal form S such that $PAQ = S$, we did not obtain P and Q. Knowing U, we can verify that $UAU^{-1} = R$. Thus the probabilistic algorithm for Rational normal form is of Las Vegas type (controllably small probability of no result), whereas the Smith normal form algorithm was of Monte Carlo type (controllably small probability of incorrect result).

To compute the transform U, first compute R via the Smith form of $xI - A$, as indicated above. Then solve the linear system $UA = RU$. An arbitrary U satisfying this equation will not do, as it may be singular. However, we may do the following. We compute a basis $U_1, \cdots, U_k$ of the solution space. Let $\lambda_1, \cdots, \lambda_k$ be indeterminants and let

$$\pi(\lambda_1, \cdots, \lambda_k) = \det\left(\sum_{i=1}^{k} \lambda_i U_i\right).$$

We choose $r_1, \cdots, r_k$ at random from F and let

$$U = \sum_{i=1}^{k} r_i U_i.$$

Then U is nonsingular unless $\pi(r_1, \cdots, r_k) = 0$. We know that π is not identically zero since if R is the rational form of A, then, by definition, a nonsingular U such that $UA = RU$ must exist. By Schwartz' result [19] the probability that we unluckily obtain a singular U is less than deg $(\pi)/s^k$, where s is the size of the set from which we choose $(r_1, \cdots, r_k)$. Thus, if det (U) is nonzero R is a verified rational canonical form of A. If it is zero then we return no solution. Either we were unlucky in computing R via the probabilistic Smith normal form algorithm, or we were unlucky in computing U. If F is finite, and a larger s is desired, the r_i may be chosen from an extension of F.

More details on this and the construction of the Jordan normal form can be found in [22].

6. Conclusion. In the meantime, we have discovered a Las Vegas solution for the Smith normal form problem of polynomial matrices [23]. This solution hinges greatly on the Hermite normal form process, as opposed to the Monte–Carlo solution proposed here. Its analysis, however, is similar to the one here. The new algorithm also finds the multipliers. In the future we will carry out practical experiments with our randomized algorithms.

REFERENCES

[1] A. Bachem and R. Kannan, *Applications of polynomial Smith normal calculations*, Report #78119-OR, Institut fur Okonometrie und Operations Research, Univ. of Bonn, Bonn, West Germany, 1978.

[2] A. Borodin, J. von zur Gathen and J. Hopcroft, *Fast parallel matrix and* GCD *computations*, Inform. Control, 52 (1982), pp. 241–256.

[3] A. Borodin, S. A. Cook and N. Pippenger, *Parallel computation for well-endowed rings and space-bounded probablistic machines*, Tech. Report 162/83, Univ. of Toronto, Toronto, Ontario, Canada, April 1983.

[4] S. A. Cook, *The classification of problems which have fast parallel algorithms*, Proc. International Conference on Foundations of Computation Theory, Borgholm, 1983, Lecture Notes in Computer Science 158, Springer–Verlag, Berlin–NewYork, pp. 78–93.

[5] I. Gohberg, P. Lancaster and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[6] A. J. Goldstein and R. L. Graham, *A Hadamard-type bound on the coefficients of a determinant of polynomials*, SIAM Rev., 16 (1974), pp. 394–395.

[7] C. Hermite, *Sur l'introduction des variables continues dans la théorie des nombres*, J. Reine Angew. Math., 41 (1851), pp. 191–216.

[8] G. Hermann, *Die Frage der endlich vielen Schritte in der Theorie der Polynomideale*, Math. Ann., 95 (1922), pp. 736–788.

[9] D. Hilbert, *Über die Theorie der algebraischen Formen*, Math. Ann., 36 (1890), pp. 473–534.

[10] O. H. Ibarra, S. Moran and L. E. Rosier, *A note on the parallel complexity of computing the rank of order n matrices*, Inform. Process. Lett., 11 (1980), p. 162.

[11] E. Kaltofen, *Effective Hilbert irreducibility*, EUROSAM 1984, Lecture Notes in Computer Science 174, Springer–Verlag, Berlin–New York, 1984, pp. 277–284.

[12] R. Kannan, *Polynomial-time algorithms for solving systems of linear equations over polynomials*, Theoret. Comput. Sci., 39 (1985), pp. 69–88.

[13] R. Kannan and A. Bachem, *Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix*, SIAM J. Comput., 8 (1979), pp. 499–507.

[14] D. E. Knuth, *The Art of Programming*, Vol. 2, *Seminumerical Algorithms*, 2nd ed., Addison–Wesley, Reading, MA, 1981.

[15] E. W. Mayr and A. R. Meyer, *The complexity of the word problems for commutative semigroups and polynomial ideals*, Adv. in Math., 46 (1982), pp. 305–329.

[16] K. Mulmuley, *A fast parallel algorithm to compute the rank of a matrix over an arbitrary field*, Combinatorica, 7 (1987), pp. 101–104.

[17] M. Newman, *Integral Matrices*, Academic Press, New York, 1972.

[18] V. Ramachandran, *Exact reduction of a polynomial matrix to the Smith normal form*, IEEE Trans. Automat. Control, AC-24 (1979), 638–641.

[19] J. T. Schwartz, *Fast probabilistic algorithms for verification of polynomial identities*, J. Assoc. Comput. Mach., 27 (1980), pp. 701–717.

[20] H. J. S. Smith, *On systems of linear indeterminate equations and congruences*, Philos. Trans. Roy. Soc. London, Ser. A, 151 (1861), pp. 293–326.

[21] B. L. van der Waerden, *Algebra*, Frederic Ungar, New York, 1970.

[22] E. Kaltofen, M. Krishnamoorthy and B. D. Saunders, *Fast parallel algorithms for similarity of matrices*, Proc. 1986 ACM Symp. Symbolic Algebraic Comput., pp. 65–70.

[23] ———, *Mr. Smith goes to Las Vegas, randomized parallel computation of the Smith normal form of polynomial matrices*, manuscript, June 1987.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

016

THE CASE OF EQUALITY IN HOPF'S INEQUALITY*

LORENZO O. HILLIARD†

Abstract. Hopf's inequality states that the subdominant eigenvalues λ of a positive n-square matrix A satisfy

$$|\lambda| \leqq \frac{M-m}{M+m} \lambda_p$$

where λ_p is the Perron eigenvalue of A and M, m are, respectively, the maximum and minimum entries of A. A complete analysis of the case of equality in Hopf's inequality is given. If A has an eigenvalue λ which satisfies the case of equality, it is shown that λ is real and the structure of the matrix A is determined.

Key words. positive matrices, stochastic matrices, eigenvalues, Hopf's inequality

AMS(MOS) subject classifications. 15A48, 15A51

1. Introduction. We shall consider properties of eigenvalues of positive matrices, that is, matrices with positive entries. The first significant work on this subject was done by Perron. His main result may be stated:

> Let A be a positive $n \times n$ matrix. Then there exists an eigenvalue $\lambda_p > 0$, which we call the Perron eigenvalue, such that λ_p has multiplicity one and such that for every other eigenvalue λ of A, $|\lambda| < \lambda_p$. The corresponding Perron eigenvector p is positive.

The theorem was later generalized by Frobenius to allow nonnegative matrices. In this case there may be other eigenvalues λ satisfying $|\lambda| = \lambda_p$. For a statement of these results, see [3].

In the case of positive integral operators, E. Hopf [1] gave a more precise form of this result. In the case of positive matrices, this result states that the subdominant eigenvalues satisfy the inequality

$$|\lambda| \leqq \frac{M-m}{M+m} \lambda_p \tag{1.1}$$

where M and m are respectively the maximum and minimum entries of the matrix. In this paper we shall determine completely the structure of a matrix A for which there is an eigenvalue λ which satisfies (1.1) with equality. We show that λ must be real and that, after a permutation, A takes a particular form. As a consequence of our results, the order of A is even and the rank of A is two, all eigenvalues of A other than λ and λ_p being zero. Since the set of positive matrices with fixed M and m form a compact set, it follows that when n is odd, there is an improvement on the inequality (1.1). It would be of interest to find the best inequality of the form (1.1) for the subdominant eigenvalues of a positive matrix.

There is considerable literature on the estimation of subdominant eigenvalues of a positive matrix. Rothblum and Tan [4] give a thorough survey of this literature. Most of the inequalities in [4] require information about the matrix A beyond the entries of A. Usually, a knowledge of the Perron eigenvector is required, and in some cases, information concerning the field of values or other quantities are required. The Hopf inequality

* Received by the editors December 29, 1986; accepted for publication May 21, 1987.

† Department of Mathematics, University of the District of Columbia, Washington, D.C. 20008.

only uses information that is easily obtained from the matrix entries, and it is of special interest for this reason.

In § 2, we give a proof of (1.1) following Ostrowski [2]. Section 3 gives our results on the case of equality. We denote by $\mathbf{R}^n$ and $\mathbf{C}^n$ the n-dimensional space of real and complex vectors, respectively. If $x \in \mathbf{R}^n$, we say that $x > 0$ if each component $x_j > 0$, and we use a similar notation for matrices. In the rest of the paper, A denotes a positive matrix of order n, with entries a_{ij}, and $M = \max_{1 \leqq i,j \leqq n} a_{ij}$ and $m = \min_{1 \leqq i,j \leqq n} a_{ij}$.

2. Hopf's inequality for positive matrices. Let $x \in \mathbf{C}^n$ and $y \in \mathbf{R}^n$, with $y > 0$. Following Hopf, we define the oscillation, osc (x/y), by

$$\operatorname{osc} \frac{x}{y} = \max_{s,t} \left| \frac{x_s}{y_s} - \frac{x_t}{y_t} \right|.$$

We show in Lemma 2.3 that

$$\operatorname{osc} \left(\frac{Ax}{Ay} \right) \leqq \frac{M-m}{M+m} \operatorname{osc} \frac{x}{y}$$

and we then use this to prove the Hopf inequality, Theorem 2.1. We start with some preliminary lemmas.

LEMMA 2.1. *If* $0 \leqq v \leqq u \leqq 1$, *then*

$$0 \leqq u - v \leqq \frac{u^{1/2}(1-v)^{1/2} - v^{1/2}(1-u)^{1/2}}{u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2}}.$$

Proof. From the arithmetic-geometric mean inequality,

$$u^{1/2}(1-v)^{1/2} \leqq \tfrac{1}{2}(u+1-v)$$

and

$$v^{1/2}(1-u)^{1/2} \leqq \tfrac{1}{2}(v+1-u).$$

When we add these inequalities,

$$u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2} \leqq 1,$$

so

$$1 \leqq \frac{1}{(u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2})^2}. \tag{2.1}$$

Multiplying both sides of (2.1) by $u - v$, we have

$$\begin{aligned} u - v &\leqq \frac{u-v}{(u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2})^2} \\ &= \frac{(u^{1/2}(1-v)^{1/2})^2 - (v^{1/2}(1-u)^{1/2})^2}{(u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2})^2} \\ &= \frac{u^{1/2}(1-v)^{1/2} - v^{1/2}(1-u)^{1/2}}{u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2}}. \end{aligned}$$

LEMMA 2.2. *Let* $p, f \in \mathbf{R}^n$ *with* $p > 0$ *and set*

$$D = \max_{1 \leqq s \leqq n} \frac{f_s}{p_s},$$

$$d = \min_{1 \leqq s \leqq n} \frac{f_s}{p_s}.$$

Let α, β be indices with $1 \leqq \alpha, \beta \leqq n$ and let $(Af)_\alpha$ denote the αth component of Af. If

$$f \text{ is not proportional to } p, \tag{2.2}$$

and if we define $K_{\alpha\beta}(f)$ by

$$K_{\alpha\beta}(f)=\left\{\frac{[A(f-dp)]_\alpha[A(Dp-f)]_\beta}{[A(Dp-f)]_\alpha[A(f-dp)]_\beta}\right\}^{1/2} \tag{2.3}$$

then

$$\left|\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}\right| \leqq \left|\frac{K_{\alpha\beta}(f)-1}{K_{\alpha\beta}(f)+1}\right| \operatorname{osc}\frac{f}{p}. \tag{2.4}$$

Proof. First we shall show that

$$\operatorname{osc}\frac{f}{p}=D-d. \tag{2.5}$$

By the definition of oscillation, we see that

$$\operatorname{osc}\frac{f}{p}\geqq D-d.$$

Let j, k be indices so that

$$\operatorname{osc}\frac{f}{p}=\left|\frac{f_j}{p_j}-\frac{f_k}{p_k}\right|.$$

But

$$\frac{f_j}{p_j}\leqq D \quad\text{and}\quad \frac{f_k}{p_k}\geqq d.$$

Thus

$$\operatorname{osc}\frac{f}{p}\leqq D-d$$

which proves (2.5). Set

$$\tilde{f}=f-dp. \tag{2.6}$$

Then

$$Dp-f=p\operatorname{osc}\frac{f}{p}-\tilde{f} \tag{2.7}$$

follows immediately from (2.5) and (2.6). Since

$$dp_s\leqq f_s\leqq Dp_s,$$

then

$$0\leqq \tilde{f}_s\leqq p_s\operatorname{osc}\frac{f}{p}$$

for $s = 1, 2, \cdots, n$. Replacing each component of $\tilde{f}$ by $p_s \operatorname{osc}(f/p)$, we have

$$\frac{(A\tilde{f})_s}{(Ap)_s}\leqq \operatorname{osc}\frac{f}{p}.$$

Thus we may define u and v, with $0 \leqq u, v \leqq 1$, by

$$\frac{(A\tilde{f})_\alpha}{(Ap)_\alpha}=u\operatorname{osc}\frac{f}{p} \tag{2.8}$$

and

$$\frac{(A\tilde{f})_\beta}{(Ap)_\beta} = v \operatorname{osc} \frac{f}{p}. \tag{2.9}$$

By equation (2.8)

$$\frac{[A(p \operatorname{osc}(f/p) - \tilde{f})]_\alpha}{(Ap)_\alpha} = \frac{(Ap)_\alpha}{(Ap)_\alpha} \operatorname{osc} \frac{f}{p} - \frac{(A\tilde{f})_\alpha}{(Ap)_\alpha}$$

$$= \operatorname{osc} \frac{f}{p} - u \operatorname{osc} \frac{f}{p} \tag{2.10}$$

$$= (1-u) \operatorname{osc} \frac{f}{p}$$

and similarly,

$$\frac{[A(p \operatorname{osc}(f/p) - \tilde{f})]_\beta}{(Ap)_\beta} = (1-v) \operatorname{osc} \frac{f}{p}. \tag{2.11}$$

From (2.2) and (2.6)–(2.11), $u \neq 1$, $v \neq 1$ and $u \neq 0$, $v \neq 0$. Thus we may assume $0 < v \leqq u < 1$. By Lemma 2.1 and (2.8)–(2.11),

$$\left| u \operatorname{osc} \frac{f}{p} - v \operatorname{osc} \frac{f}{p} \right| \leqq \left| \frac{u^{1/2}(1-v)^{1/2} - v^{1/2}(1-u)^{1/2}}{u^{1/2}(1-v)^{1/2} + v^{1/2}(1-u)^{1/2}} \right| \operatorname{osc} \frac{f}{p}$$

$$= \left| \frac{u^{1/2}(1-v)^{1/2} \operatorname{osc}(f/p) - v^{1/2}(1-u)^{1/2} \operatorname{osc}(f/p)}{u^{1/2}(1-v)^{1/2} \operatorname{osc}(f/p) + v^{1/2}(1-u)^{1/2} \operatorname{osc}(f/p)} \right| \operatorname{osc} \frac{f}{p}.$$

Then

$$\left| \frac{(A\tilde{f})_\alpha}{(Ap)_\alpha} - \frac{(A\tilde{f})_\beta}{(Ap)_\beta} \right| \leqq \frac{X-Y}{X+Y} \operatorname{osc} \frac{f}{p}$$

where

$$X = \left\{ \frac{(A\tilde{f})_\alpha}{(Ap)_\alpha} \frac{[A(p \operatorname{osc}(f/p) - \tilde{f})]_\beta}{(Ap)_\beta} \right\}^{1/2},$$

$$Y = \left\{ \frac{[A(p \operatorname{osc}(f/p) - \tilde{f})]_\alpha}{(Ap)_\alpha} \frac{(A\tilde{f})_\beta}{(Ap)_\beta} \right\}^{1/2}.$$

Since

$$X = \left\{ \frac{[A(f-dp)]_\alpha}{(Ap)_\alpha} \frac{[A(Dp-f)]_\beta}{(Ap)_\beta} \right\}^{1/2},$$

$$Y = \left\{ \frac{[A(Dp-f)]_\alpha}{(Ap)_\alpha} \frac{[A(f-dp)]_\beta}{(Ap)_\beta} \right\}^{1/2},$$

we have

$$K_{\alpha\beta}(f) = X/Y,$$

and hence

$$\left| \frac{(A\tilde{f})_\alpha}{(Ap)_\alpha} - \frac{(A\tilde{f})_\beta}{(Ap)_\beta} \right| \leqq \left| \frac{K_{\alpha\beta}(f) - 1}{K_{\alpha\beta}(f) + 1} \right| \operatorname{osc} \frac{f}{p}.$$

To show (2.4), we write

$$\begin{aligned}\left|\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}\right| &= \left|\frac{[A(f-dp)]_\alpha}{(Ap)_\alpha}-\frac{[A(f-dp)]_\beta}{(Ap)_\beta}\right| \\ &= \left|\frac{(A\tilde{f})_\alpha}{(Ap)_\alpha}-\frac{(A\tilde{f})_\beta}{(Ap)_\beta}\right| \\ &\leqq \left|\frac{K_{\alpha\beta}(f)-1}{K_{\alpha\beta}(f)+1}\right| \operatorname{osc}\frac{f}{p}.\end{aligned}$$

Therefore (2.4) holds and the proof is finished.

LEMMA 2.3. (a) *If* f, $p \in \mathbf{R}^n$, *with* $p > 0$, *and* (2.2) *holds, then*

$$1 \leqq K_{\alpha\beta}(f) \leqq \frac{M}{m}, \qquad 1 \leqq \alpha, \beta \leqq n. \tag{2.12}$$

(b) *If* $f \in \mathbf{C}^n$, $p \in \mathbf{R}^n$ *with* $p > 0$, *and* (2.2) *holds, then*

$$\left|\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}\right| \leqq \frac{M-m}{M+m}\operatorname{osc}\frac{f}{p}, \qquad 1 \leqq \alpha, \beta \leqq n. \tag{2.13}$$

Proof. (a) Let g, $h \in \mathbf{R}^n$ be nonnegative vectors. Consider

$$\begin{aligned}\frac{(Ah)_\alpha (Ag)_\beta}{(Ag)_\alpha (Ah)_\beta} &= \frac{\sum_{l=1}^n a_{\alpha l} h_l \sum_{l=1}^n a_{\beta l} g_l}{\sum_{l=1}^n a_{\alpha l} g_l \sum_{l=1}^n a_{\beta l} h_l} \\ &\leqq \frac{\sum_{l=1}^n M h_l \sum_{l=1}^n M g_l}{\sum_{l=1}^n m g_l \sum_{l=1}^n m h_l} \\ &= \frac{M^2}{m^2}.\end{aligned} \tag{2.14}$$

Setting $h = f - dp$, $g = Dp - f$, we obtain from the definition of $K_{\alpha\beta}$

$$K_{\alpha\beta}(f) \leqq \frac{M}{m},$$

which shows one of the inequalities of (2.12). To show $K_{\alpha\beta}(f) \geqq 1$ we use the definition of $K_{\alpha\beta}(f)$ and (2.5)–(2.9) as follows:

$$\begin{aligned}K_{\alpha\beta}(f) &= \left\{\frac{[A(f-dp)]_\alpha [A(Dp-f)]_\beta}{[A(Dp-f)]_\alpha [A(f-dp)]_\beta}\right\}^{1/2} \\ &= \left\{\frac{(A\tilde{f})_\alpha}{(Ap)_\alpha}\frac{[A(p\operatorname{osc}(f/p)-\tilde{f})]_\beta}{(Ap)_\beta} \Big/ \frac{[A(p\operatorname{osc}(f/p)-\tilde{f})]_\alpha}{(Ap)_\alpha}\frac{(A\tilde{f})_\beta}{(Ap)_\beta}\right\}^{1/2} \\ &= \left\{\frac{u\operatorname{osc}(f/p)(1-v)\operatorname{osc}(f/p)}{(1-u)\operatorname{osc}(f/p)v\operatorname{osc}(f/p)}\right\}^{1/2} \\ &= \left\{\frac{u}{v}\frac{1-v}{1-u}\right\}^{1/2}.\end{aligned} \tag{2.15}$$

Since $u \geqq v$ and $1 - v \geqq 1 - u$, we have $K_{\alpha\beta}(f) \geqq 1$. Therefore part (a) of the lemma is true.

(b) Equation (2.13) obviously holds if

$$\frac{(Af)_\alpha}{(Ap)_\alpha}=\frac{(Af)_\beta}{(Ap)_\beta}.$$

To show (2.13) for

$$\frac{(Af)_\alpha}{(Ap)_\alpha}\neq\frac{(Af)_\beta}{(Ap)_\beta}$$

we first suppose that $f\in \mathbf{R}^n$. Note that $(K-1)/(K+1)$ is a monotonically increasing function of K. From part (a),

$$\begin{aligned}\frac{K_{\alpha\beta}(f)-1}{K_{\alpha\beta}(f)+1}&\leqq\frac{M/m-1}{M/m+1}\\&=\frac{M-m}{M+m}.\end{aligned}\tag{2.16}$$

Therefore, by Lemma 2.2 and the inequality above,

$$\left|\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}\right|\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{f}{p}.$$

Hence (2.13) is true for real f.

Suppose $f\in \mathbf{C}^n$ and write

$$\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}=\eta t$$

where η is a complex number with $|\eta|=1$ and $t>0$. Then

$$\frac{(A\bar\eta f)_\alpha}{(Ap)_\alpha}-\frac{(A\bar\eta f)_\beta}{(Ap)_\beta}=t.$$

Obviously,

$$\operatorname{osc}\frac{\bar\eta f}{p}=\operatorname{osc}\frac{f}{p}.$$

Then if we apply (2.13) to the real vector Re $(\bar\eta f)$,

$$\begin{aligned}t&=\operatorname{Re}\left|\frac{(A\bar\eta f)_\alpha}{(Ap)_\alpha}-\frac{(A\bar\eta f)_\beta}{(Ap)_\beta}\right|\\&=\frac{[A(\operatorname{Re}(\bar\eta f))]_\alpha}{(Ap)_\alpha}-\frac{[A(\operatorname{Re}(\bar\eta f))]_\beta}{(Ap)_\beta}\\&\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{\operatorname{Re}(\bar\eta f)}{p}\\&\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{\bar\eta f}{p}.\end{aligned}$$

Therefore (2.13) is true for complex vectors and the proof of the lemma is finished.

THEOREM 2.1. *Let A be a positive matrix with Perron eigenvalue λ_p. Let λ be any other eigenvalue of A. Then*

$$|\lambda|\leqq\frac{M-m}{M+m}\lambda_p.$$

Proof. Let $p > 0$ be the Perron eigenvector of A, and let f be an eigenvector corresponding to λ. By Lemma 2.3

$$\left|\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}\right| \leqq \frac{M-m}{M+m}\operatorname{osc}\frac{f}{p}.$$

Since λ, λ_p are eigenvalues of A

$$\begin{aligned}\frac{|\lambda|}{\lambda_p}\left|\frac{f_\alpha}{p_\alpha}-\frac{f_\beta}{p_\beta}\right| &= \left|\frac{\lambda f_\alpha}{\lambda_p p_\alpha}-\frac{\lambda f_\beta}{\lambda_p p_\beta}\right| \\ &= \left|\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}\right| \\ &\leqq \frac{M-m}{M+m}\operatorname{osc}\frac{f}{p}.\end{aligned}$$

Hence

$$\frac{|\lambda|}{\lambda_p}\operatorname{osc}\frac{f}{p} \leqq \frac{M-m}{M+m}\operatorname{osc}\frac{f}{p}.$$

Therefore

$$|\lambda| \leqq \frac{M-m}{M+m}\lambda_p$$

and the theorem is proved.

3. The case of equality. In this section A denotes an $n \times n$ strictly positive matrix with Perron eigenvalue λ_p, and with another eigenvalue λ satisfying

$$|\lambda| = \frac{M-m}{M+m}\lambda_p \tag{3.1}$$

where $M = \max_{1 \leqq s,t \leqq n} a_{st}$ and $m = \min_{1 \leqq s,t \leqq n} a_{st}$.

Our goal is to develop the consequences for A of the equation (3.1). The first series of lemmas leads to the result that λ must be real (Lemma 3.9). Following this, we provide a series of lemmas that lead to the final result (Theorem 3.1).

We let $p > 0$ denote a Perron eigenvector of A, and we let f denote an eigenvector of A corresponding to the eigenvalue λ. We shall suppose that $\lambda \neq 0$. If $\lambda = 0$, then $M = m$ and the matrix A consists of all entries M, so the structure is trivial in this case. We use the notation $\lambda = \lambda' + i\lambda''$ to represent λ in terms of its real and imaginary parts. Similarly, we write $f = f' + if''$, where the real vectors f', f'' are the real and imaginary parts of f. Let

$$D' = \max_{1 \leqq s \leqq n}\frac{f'_s}{p_s}, \qquad d' = \min_{1 \leqq s \leqq n}\frac{f'_s}{p_s},$$

$$D'' = \max_{1 \leqq s \leqq n}\frac{f''_s}{p_s}, \qquad d'' = \min_{1 \leqq s \leqq n}\frac{f''_s}{p_s}$$

where f'_s, f''_s are components of the vectors f', f'', respectively. Let α, β be integers with $1 \leqq \alpha$, $\beta \leqq n$ and let η, $|\eta| = 1$, be a complex number chosen so that

$$\frac{(A\eta f)_\alpha}{(Ap)_\alpha}-\frac{(A\eta f)_\beta}{(Ap)_\beta} > 0.$$

Since ηf is also an eigenvector, we may replace ηf by f, and assume that

$$\frac{(Af)_\alpha}{(Ap)_\alpha}-\frac{(Af)_\beta}{(Ap)_\beta}>0.$$

We shall say that the eigenvector f is normalized with respect to the indices α, β if f satisfies this inequality.

LEMMA 3.1. *Let λ be an eigenvalue of A satisfying* (3.1). *Then there exists an eigenvector f corresponding to λ, and indices j, k so that*

$$\text{(a)}\quad \frac{(Af)_j}{(Ap)_j}-\frac{(Af)_k}{(Ap)_k}=\frac{M-m}{M+m}\left|\frac{f_j}{p_j}-\frac{f_k}{p_k}\right|,$$

$$\text{(b)}\quad \frac{(Af')_j}{(Ap)_j}-\frac{(Af')_k}{(Ap)_k}=\frac{M-m}{M+m}\max_{1\leqq r,s\leqq n}\left|\frac{f'_r}{p_r}-\frac{f'_s}{p_s}\right|,$$

$$\text{(c)}\quad \frac{(Af)_j}{(Ap)_j}-\frac{(Af)_k}{(Ap)_k}=\frac{M-m}{M+m}\max_{1\leqq r,s\leqq n}\left|\frac{f'_r}{p_r}-\frac{f'_s}{p_s}\right|.$$

Proof. Let f be an eigenvector of A corresponding to eigenvalue λ and let j, k be indices with $1\leqq j, k\leqq n$ so that

$$\operatorname{osc}\frac{f}{p}=\left|\frac{f_j}{p_j}-\frac{f_k}{p_k}\right|.$$

Then

$$\frac{|\lambda|}{\lambda_p}\left|\frac{f_j}{p_j}-\frac{f_k}{p_k}\right|=\frac{M-m}{M+m}\left|\frac{f_j}{p_j}-\frac{f_k}{p_k}\right|.$$

Rewriting the left side of the above equation, we have

$$\left|\frac{(Af)_j}{(Ap)_j}-\frac{(Af)_k}{(Ap)_k}\right|=\frac{M-m}{M+m}\left|\frac{f_j}{p_j}-\frac{f_k}{p_k}\right|.$$

By normalizing f with respect to indices j, k, we obtain part (a) of the lemma. Also, from Lemma 2.3(b)

$$\begin{aligned}\frac{(Af)_j}{(Ap)_j}-\frac{(Af)_k}{(Ap)_k}&=\frac{(Af')_j}{(Ap)_j}-\frac{(Af')_k}{(Ap)_k}\\ &\leqq\frac{M-m}{M+m}\max_{1\leqq r,s\leqq n}\left|\frac{f'_r}{p_r}-\frac{f'_s}{p_s}\right|\\ &\leqq\frac{M-m}{M+m}\max_{1\leqq r,s\leqq n}\left|\frac{f_r}{p_r}-\frac{f_s}{p_s}\right|.\end{aligned}$$

By part (a) of the lemma, the inequalities above become equalities and the lemma is proven.

LEMMA 3.2. *Let g be a real vector not proportional to p and let*

$$d=\min_{1\leqq l\leqq n}\frac{g_l}{p_l}$$

and

$$D=\max_{1\leqq l\leqq n}\frac{g_l}{p_l}.$$

Suppose there are indices α, β *with* $1 \leqq \alpha, \beta \leqq n$ *so that*

$$K_{\alpha\beta}(g)=\frac{M}{m}.$$

Then for each $l = 1, 2, \cdots, n$,

$$a_{\alpha l}=M \quad \textit{or} \quad g_l=dp_l, \tag{3.2a}$$

$$a_{\alpha l}=m \quad \textit{or} \quad g_l=Dp_l, \tag{3.2b}$$

$$a_{\beta l}=M \quad \textit{or} \quad g_l=Dp_l, \tag{3.2c}$$

$$a_{\beta l}=m \quad \textit{or} \quad g_l=dp_l. \tag{3.2d}$$

Proof. By definition

$$K^2_{\alpha\beta}(g)=\frac{[A(g-dp)]_\alpha[A(Dp-g)]_\beta}{[A(Dp-g)]_\alpha[A(g-dp)]_\beta}$$

and we have

$$K^2_{\alpha\beta}(g)=\frac{\sum_{l=1}^n a_{\alpha l}(g_l-dp_l)\sum_{l=1}^n a_{\beta l}(Dp_l-g_l)}{\sum_{l=1}^n a_{\alpha l}(Dp_l-g_l)\sum_{l=1}^n a_{\beta l}(g_l-dp_l)} \tag{3.3}$$

$$=\frac{\sum_{l=1}^n M(g_l-dp_l)\sum_{l=1}^n M(Dp_l-g_l)}{\sum_{l=1}^n m(Dp_l-g_l)\sum_{l=1}^n m(g_l-dp_l)}. \tag{3.4}$$

The numerator of (3.3) is less than or equal to the numerator of (3.4) and similarly, the denominator of (3.3) is greater than or equal to the denominator of (3.4). Hence, since we have equality in (3.3) and (3.4),

$$\left[\sum_{l=1}^n a_{\alpha l}(g_l-dp_l)\right]\left[\sum_{l=1}^n a_{\beta l}(Dp_l-g_l)\right]=\left[\sum_{l=1}^n M(g_l-dp_l)\right]\left[\sum_{l=1}^n M(Dp_l-g_l)\right] \tag{3.5}$$

and

$$\left[\sum_{l=1}^n a_{\alpha l}(Dp_l-g_l)\right]\left[\sum_{l=1}^n a_{\beta l}(g_l-dp_l)\right]=\left[\sum_{l=1}^n m(Dp_l-g_l)\right]\left[\sum_{l=1}^n m(g_l-dp_l)\right]. \tag{3.6}$$

Note that the eight sums in (3.5) and (3.6) are positive, and that each term in each sum is nonnegative. Since $m \leqq a_{\alpha l} \leqq M$, $m \leqq a_{\beta l} \leqq M$, we conclude that

$$a_{\alpha l}(g_l-dp_l)=M(g_l-dp_l),$$

$$a_{\beta l}(Dp_l-g_l)=M(Dp-g_l),$$

$$a_{\alpha l}(Dp_l-g_l)=m(Dp_l-g_l),$$

$$a_{\beta l}(g_l-dp_l)=m(g_l-dp_l)$$

or

$$(a_{\alpha l}-M)(g_l-dp_l)=0,$$

$$(a_{\alpha l}-m)(Dp_l-g_l)=0,$$

$$(a_{\beta l}-M)(Dp_l-g_l)=0,$$

$$(a_{\beta l}-m)(g_l-dp_l)=0.$$

Thus, the lemma is true.

LEMMA 3.3. *Let $g \in \mathbf{R}^n$ and α, β be indices satisfying Lemma 3.2. Then for each $l = 1, 2, \cdots, n$ either*

(a) $a_{\alpha l} = m, \quad a_{\beta l} = M, \quad g_l = dp_l$

or

(b) $a_{\alpha l} = M, \quad a_{\beta l} = m, \quad g_l = Dp_l$.

Proof. Pick an index l and suppose that $m < a_{\alpha l} < M$. Then from (3.2a), $g_l = dp_l$, and from (3.2b), $g_l = Dp_l$, which is a contradiction. Hence either $a_{\alpha l} = m$ or $a_{\alpha l} = M$. If $a_{\alpha l} = m$, then by (3.2a), $g_l = dp_l$ so from (3.2c), $a_{\beta l} = M$, and case (a) holds. In a similar way, if $a_{\alpha l} = M$, we deduce case (b), so the lemma is true.

LEMMA 3.4. *Let f be normalized with respect to the indices j, k. Then*

(a) $K_{jk}(f') = M/m$.

(b) *For each $l = 1, 2, \cdots, n$ either*

$$a_{jl} = M, \quad a_{kl} = m, \quad f'_l = D'p_l, \quad \text{or}$$

$$a_{jl} = m, \quad a_{kl} = M, \quad f'_l = d'p_l.$$

Proof. It is easily seen that f' is not proportional to p, so $K_{jk}(f')$ is defined. By Lemma 2.3

$$K_{jk}(f') \leqq \frac{M}{m}$$

and

$$\frac{K_{jk}(f')-1}{K_{jk}(f')+1} \leqq \frac{M-m}{M+m}.$$

Then by Lemmas 2.2 and 3.1

$$\begin{aligned}\frac{(Af)_j}{(Ap)_j} - \frac{(Af)_k}{(Ap)_k} &= \frac{(Af')_j}{(Ap)_j} - \frac{(Af')_k}{(Ap)_k}\\ &\leqq \frac{K_{jk}(f')-1}{K_{jk}(f')+1} \operatorname{osc} \frac{f'}{p}\\ &\leqq \frac{M-m}{M+m} \operatorname{osc} \frac{f'}{p}\\ &= \frac{M-m}{M+m} \left| \frac{f_j}{p_j} - \frac{f_k}{p_k} \right|.\end{aligned}$$

Thus from (3.1),

$$K_{jk}(f') = \frac{M}{m}$$

and part (a) of Lemma 3.4 is true. Part (b) follows from Lemma 3.3.

Using Lemma 3.4, we define a partition $\mathscr{J}$, $\mathscr{K}$ of the set $\{1, 2, \cdots, n\}$ so that for each $l \in \mathscr{J}$,

$$a_{jl} = M, \quad a_{kl} = m, \quad f'_l = D'p_l \tag{3.7}$$

and for each $l \in \mathscr{K}$

$$a_{jl} = m, \quad a_{kl} = M, \quad f'_l = d'p_l. \tag{3.8}$$

This partition will be used in the subsequent development.

LEMMA 3.5. *Let A be an $n \times n$ positive matrix with eigenvalue λ satisfying* (3.1). *Let f be normalized with respect to indices j, k. Then*

(a) λ *is real iff* $f'_j/p_j \neq f'_k/p_k$;

(b) λ *is purely imaginary iff* $f'_j/p_j = f'_k/p_k$.

Proof. For simplification, set

$$\gamma = \operatorname{Im} \frac{(Af)_s}{(Ap)_s}, \qquad \rho_s = \operatorname{Re} \frac{(Af)_s}{(Ap)_s}$$

where $s = j$ or k. (From the normalization of f, γ does not depend on j, k.) So

$$\frac{(Af)_s}{(Ap)_s} = \frac{\lambda}{\lambda_p} \frac{f_s}{p_s} = \rho_s + i\gamma, \qquad s = j, k.$$

Solving these equations for f_j, f_k, we obtain

$$\frac{f'_s}{p_s} = \frac{\rho_s \lambda' + \gamma \lambda''}{|\lambda|^2} \lambda p, \qquad s = j, k$$

and

$$\frac{f''_s}{p_s} = \frac{\lambda' \gamma - \rho_s \lambda''}{|\lambda|^2} \lambda p, \qquad s = j, k.$$

Case 1. Suppose $f'_j/p_k \neq f'_k/p_k$. Then

$$\frac{(Af)_j}{(Ap)_j} - \frac{(Af)_k}{(Ap)_k} = \frac{M-m}{M+m} \max \left| \frac{f'_r}{p_r} - \frac{f'_s}{p_s} \right|$$

$$= \frac{M-m}{M+m} (D' - d')$$

$$= \frac{M-m}{M+m} \left| \frac{f'_j}{p_j} - \frac{f'_k}{p_k} \right|.$$

Hence

$$\rho_j - \rho_k = \frac{M-m}{M+m} \left| \frac{\rho_j \lambda' + \gamma \lambda''}{|\lambda|^2} \lambda_p - \frac{\rho_k \lambda' + \gamma \lambda''}{|\lambda|^2} \lambda_p \right|$$

$$= \frac{M-m}{M+m} \left| \frac{(\rho_j - \rho_k)\lambda'}{|\lambda|^2} \right| \lambda_p.$$

Since $p_j > \rho_k$,

$$1 = \frac{|\lambda'|}{|\lambda|}.$$

Therefore,

$$|\lambda| = |\lambda'|$$

and λ is real.

Case 2. Assume $f'_j/p_j = f'_k/p_k$. By the normalization of f,

$$\frac{(Af)_j}{(Ap)_j} - \frac{(Af)_k}{(Ap)_k} = \frac{M-m}{M+m} \left| \frac{f_j}{p_j} - \frac{f_k}{p_k} \right|.$$

Hence

$$\rho_j - \rho_k = \frac{|\lambda|}{\lambda_p} \left| \frac{\lambda' \gamma - \rho_j \lambda''}{|\lambda|^2} \lambda_p - \frac{\lambda' \gamma - \rho_k \lambda''}{|\lambda|^2} \lambda_p \right|$$

$$= \frac{|\lambda|}{\lambda_p |\lambda|^2} \left| -\rho_j \lambda'' + \rho_k \lambda'' \right| \lambda_p.$$

Then

$$|\lambda|(\rho_j - \rho_k) = |\lambda''| |-\rho_j + \rho_k|$$

or equivalently

$$|\lambda|=|\lambda''|.$$

Thus λ is pure imaginary.

LEMMA 3.6. *Suppose λ is pure imaginary. Then*

$$\operatorname{osc}\frac{f}{p}=\operatorname{osc}\frac{f''}{p}=\operatorname{osc}\frac{f'}{p}.$$

Proof. By Lemma 2.3 applied to the vector f',

$$\frac{|\lambda''|}{\lambda_p}\left|\frac{f''_\alpha}{p_\alpha}-\frac{f''_\beta}{p_\beta}\right|\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{f'}{p}.$$

Hence, taking the maximum over α, β,

$$\frac{|\lambda''|}{\lambda_p}\operatorname{osc}\frac{f''}{p}\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{f'}{p}.$$

Thus,

$$\operatorname{osc}\frac{f''}{p}\leqq\operatorname{osc}\frac{f'}{p}.$$

Using similar arguments we can show that the inequality

$$\left|\frac{(Af'')_\alpha}{(Ap)_\alpha}-\frac{(Af'')_\beta}{(Ap)_\beta}\right|\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{f''}{p}$$

reduces to

$$\operatorname{osc}\frac{f'}{p}\leqq\operatorname{osc}\frac{f''}{p}.$$

Note that $\operatorname{osc}(f/p)=\operatorname{osc}(f'/p)$ by Lemma 3.1(a), (c). Thus the lemma is true.

We now define indices u and v, with $1\leqq u, v\leqq n$, by

$$\operatorname{osc}\frac{f'}{p}=\left|\frac{f'_u}{p_u}-\frac{f'_v}{p_v}\right|. \tag{3.9}$$

We have the following lemma.

LEMMA 3.7. *Let A be an $n\times n$ positive matrix with eigenvalue $\lambda=i\lambda''$ satisfying* (3.1) *and let f be the corresponding eigenvector, normalized with respect to j, k. Then*

(a) $K_{uv}(f'')=M/m$,

(b) *For each $l=1, 2, \cdots, n$ either*

$$a_{ul}=M,\quad a_{vl}=m,\quad f''_l=D''p_l,\quad \textit{or}$$

$$a_{ul}=m,\quad a_{vl}=M,\quad f''_l=d''p_l.$$

Proof. It is easily seen that f' is not proportional to p, so $K_{uv}(f'')$ is defined. From Lemmas 2.2 and 3.6,

$$\begin{aligned}\frac{|\lambda''|}{\lambda_p}\left|\frac{f'_u}{p_u}-\frac{f'_v}{p_v}\right|&=\left|\frac{(Af'')_u}{(Ap)_u}-\frac{(Af'')_v}{(Ap)_v}\right|\\&\leqq\frac{K_{uv}(f'')-1}{K_{uv}(f'')+1}\operatorname{osc}\frac{f''}{p}\\&\leqq\frac{M-m}{M+m}\operatorname{osc}\frac{f'}{p}.\end{aligned}\tag{3.10}$$

By the definition of λ, u, v we must have equality in inequalities of (3.10). Thus

$$K_{uv}(f'') = \frac{M}{m}.$$

Part (b) of the lemma follows from part (a) and Lemma 3.3.

Using Lemma 3.7, we may define a partition $\mathscr{U}$, $\mathscr{V}$ of the set $\{1, 2, \cdots, n\}$ so that

$$a_{ul} = M, \quad a_{vl} = m, \quad f''_l = D''p_l \quad \text{for each } l \in \mathscr{U}, \tag{3.11}$$

$$a_{ul} = m, \quad a_{vl} = M, \quad f''_l = d''p_l \quad \text{for each } l \in \mathscr{V}. \tag{3.12}$$

LEMMA 3.8. *We have*

(a) $f''_j + f''_k = f''_u + f''_v$,

(b) $p_j + p_k = p_u + p_v$.

Proof. Let $\mathscr{J}$, $\mathscr{K}$ and $\mathscr{U}$, $\mathscr{V}$ be partitions on the set $\{1, 2, \cdots, n\}$ as defined in equations (3.7), (3.8), (3.11) and (3.12), respectively. The real parts of rows j, k, u, v of the matrix equation $Af = \lambda f$ are given by

$$\sum_{l \in \mathscr{J} \cap \mathscr{U}} a_{jl}f'_l + \sum_{l \in \mathscr{J} \cap \mathscr{V}} a_{jl}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{U}} a_{jl}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{V}} a_{jl}f'_l = -\lambda''f''_j,$$

$$\sum_{l \in \mathscr{J} \cap \mathscr{U}} a_{kl}f' + \sum_{l \in \mathscr{J} \cap \mathscr{V}} a_{kl}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{U}} a_{kl}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{V}} a_{kl}f'_l = -\lambda''f''_k,$$

$$\sum_{l \in \mathscr{J} \cap \mathscr{U}} a_{ul}f'_l + \sum_{l \in \mathscr{J} \cap \mathscr{V}} a_{ul}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{U}} a_{ul}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{V}} a_{ul}f'_l = -\lambda''f''_u,$$

$$\sum_{l \in \mathscr{J} \cap \mathscr{U}} a_{vl}f'_l + \sum_{l \in \mathscr{J} \cap \mathscr{V}} a_{vl}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{U}} a_{vl}f'_l + \sum_{l \in \mathscr{K} \cap \mathscr{V}} a_{vl}f'_l = -\lambda''f''_v.$$

Using the definition of the partitioned sets $\mathscr{J}$, $\mathscr{K}$, $\mathscr{U}$, $\mathscr{V}$ we obtain

$$MD' \sum_{l \in \mathscr{J} \cap \mathscr{U}} p_l + MD' \sum_{l \in \mathscr{J} \cap \mathscr{V}} p_l + md' \sum_{l \in \mathscr{K} \cap \mathscr{U}} p_l + md' \sum_{l \in \mathscr{K} \cap \mathscr{V}} p_l = -\lambda''f_j, \tag{3.13}$$

$$mD' \sum_{l \in \mathscr{J} \cap \mathscr{U}} p_l + mD' \sum_{l \in \mathscr{J} \cap \mathscr{V}} p_l + Md' \sum_{l \in \mathscr{K} \cap \mathscr{U}} p_l + Md' \sum_{l \in \mathscr{K} \cap \mathscr{V}} p_l = -\lambda''f''_k, \tag{3.14}$$

$$MD' \sum_{l \in \mathscr{J} \cap \mathscr{U}} p_l + mD' \sum_{l \in \mathscr{J} \cap \mathscr{V}} p_l + Md' \sum_{l \in \mathscr{K} \cap \mathscr{U}} p_l + md' \sum_{l \in \mathscr{K} \cap \mathscr{V}} p_l = -\lambda''f''_u, \tag{3.15}$$

$$mD' \sum_{l \in \mathscr{J} \cap \mathscr{U}} p_l + MD' \sum_{l \in \mathscr{J} \cap \mathscr{V}} p_l + md' \sum_{l \in \mathscr{K} \cap \mathscr{U}} p_l + Md' \sum_{l \in \mathscr{K} \cap \mathscr{V}} p_l = -\lambda''f''_v. \tag{3.16}$$

Adding (3.13) to (3.14) and (3.15) to (3.16), we obtain part (a) of the lemma.

Similarly, part (b) of the lemma may be proven using rows j, k, u, v of the matrix equation $Ap = \lambda_p p$.

LEMMA 3.9. *Let A be an $n \times n$ positive matrix with eigenvalue λ satisfying* (3.1). *Then λ is real.*

Proof. Suppose

$$\frac{f'_j}{p_j} = \frac{f'_k}{p_k}.$$

By Lemma 3.5, λ is pure imaginary. Without loss of generality, we may take

$$\frac{f''_j}{p_j} = D'' \tag{3.17}$$

and

$$\frac{f''_k}{p_k} = d''. \tag{3.18}$$

By the definition of u, v of (3.9) and Lemma 3.6

$$\frac{f''_u}{p_u} = \frac{f''_v}{p_v}$$

and

$$\left\{\frac{f'_u}{p_u}, \frac{f'_v}{p_v}\right\} = \{D', d'\}.$$

Suppose $f''_u = d''p_u$ and $f''_v = d''p_v$. Then by Lemma 3.8 and (3.17) and (3.18)

$$\begin{aligned} D''p_j + d''p_k &= d''p_u + d''p_v \\ &= d''(p_u + p_v) \\ &= d''p_j + d''p_k. \end{aligned}$$

Then

$$D''p_j = d''p_j$$

or

$$D'' = d'',$$

which is a contradiction by Lemma 3.6. We may obtain similar results if we set

$$\frac{f''_u}{p_u} = D'' \quad \text{and} \quad \frac{f''_v}{p_v} = D''$$

and

$$\frac{f''_j}{p_j} = d'' \quad \text{and} \quad \frac{f''_k}{p_k} = D''.$$

Hence, the lemma is true. □

In the next series of lemmas we assume that λ is a real eigenvalue satisfying (3.1) and we determine the structure of A. We write $D'' = D$ and $d' = d$, so that Lemma 3.4 becomes

$$a_{jl} = M,\ a_{kl} = m,\ f_l = Dp_l \quad \text{for each } l \in \mathscr{J}$$

and

$$a_{jl} = m,\ a_{kl} = M,\ f_l = dp_l \quad \text{for each } l \in \mathscr{K}.$$

LEMMA 3.10. *Let f be normalized with respect to indices j, k.*

(a) *If $\lambda > 0$, then*

$$f_j = Dp_j \quad \textit{and} \quad f_k = dp_k.$$

(b) *If $\lambda < 0$, then*

$$f_j = dp_j \quad \textit{and} \quad f_k = Dp_k.$$

Proof. Assume $\lambda > 0$. Then by assumption on j, k

$$\frac{(Af)_j}{(Ap)_j} > \frac{(Af)_k}{(Ap)_k}$$

if and only if

$$\frac{f_j}{p_j} > \frac{f_k}{p_k}$$

if and only if

$$\frac{f_j}{p_j} = D \quad \text{and} \quad \frac{f_k}{p_k} = d.$$

Hence part (a) of the lemma is true. By a similar proof part (b) can be shown to be true.

LEMMA 3.11. *Let A_l, $l = 1, 2, \cdots, n$, be the row vectors of A. If $\lambda > 0$, then*

$$A_l = \begin{cases} A_j & \text{if } l \in \mathscr{J}, \\ A_k & \text{if } l \in \mathscr{K} \end{cases}$$

and if $\lambda < 0$, then

$$A_l = \begin{cases} A_k & \text{if } l \in \mathscr{J}, \\ A_j & \text{if } l \in \mathscr{K}. \end{cases}$$

Proof. Let f be normalized with respect to indices j, k and suppose $\lambda > 0$. Let $t \in \mathscr{J}$. Then $f_t = Dp_t$ and

$$\begin{aligned} \frac{(Af)_t}{(Ap)_t} - \frac{(Af)_k}{(Ap)_k} &= \frac{\lambda f_t}{\lambda_p p_t} - \frac{\lambda f_k}{\lambda_p p_k} \\ &= \frac{\lambda}{\lambda_p}(D-d) \\ &> 0. \end{aligned}$$

We see that f is also normalized with respect to the indices t, k. By Lemma 3.4

$$a_{tl} = M, \quad a_{kl} = m, \quad f_l = Dp_l \quad \text{if } l \in \mathscr{J}$$

and

$$a_{tl} = m, \quad a_{kl} = M, \quad f_l = dp_l \quad \text{if } l \in \mathscr{K}.$$

Therefore,

$$A_t = A_j.$$

Using a similar argument, we may show that for each $s \in \mathscr{K}$

$$a_{jl} = M, \quad a_{sl} = m, \quad f_l = Dp_l \quad \text{if } l \in \mathscr{J}$$

and

$$a_{jl} = m, \quad a_{sl} = M, \quad f_l = dp_l \quad \text{if } l \in \mathscr{K}.$$

Hence $A_s = A_k$. The proof for the case $\lambda < 0$ is similar and thus the lemma is true.

LEMMA 3.12. *Let*

$$\Phi_1 = \sum_{l \in \mathscr{J}} p_l$$

and

$$\Phi_2 = \sum_{l \in \mathscr{K}} p_l.$$

Then

$$\Phi_1 = \Phi_2.$$

Proof. The equations

$$(Af)_j = \lambda_j f_j \quad \text{and} \quad (Af)_k = \lambda f_k$$

are equivalent to

(3.19) $$\sum_{l \in \mathscr{J}} a_{jl} f_l + \sum_{l \in \mathscr{K}} a_{jl} f_l = \lambda f_j$$

and

$$\sum_{l\in\mathcal{J}} a_{kl}f_l + \sum_{l\in\mathcal{K}} a_{kl}f_l = \lambda f_k. \tag{3.20}$$

Without loss of generality, we may assume $\lambda > 0$. If we use the definitions for $\mathcal{J}$, $\mathcal{K}$, Φ_1, Φ_2 and Lemma 3.10, then (3.19) and (3.20) are equivalent to

$$MD\Phi_1 + md\Phi_2 = \lambda Dp_j, \tag{3.21}$$

$$mD\Phi_1 + Md\Phi_2 = \lambda dp_k. \tag{3.22}$$

Multiplying (3.21) by p_k and (3.22) by p_j and subtracting the result, we have

$$(MD\Phi_1 + md\Phi_2)p_k - (mD\Phi_1 + Md\Phi_2)p_j = \lambda p_jp_k(D-d). \tag{3.23}$$

To obtain an expression for p_j and p_k, we consider the equations

$$(Ap)_j = \lambda_p p_j \quad\text{and}\quad (Ap)_k = \lambda_p p_k,$$

which are equivalent to

$$M\Phi_1 + m\Phi_2 = \lambda_p p_j, \qquad m\Phi_1 + M\Phi_2 = \lambda_p p_k.$$

Then

$$\begin{aligned}(MD\Phi_1 + md\Phi_2)p_k &= (MD\Phi_1 + md\Phi_2)(m\Phi_1 + M\Phi_2)\frac{1}{\lambda_p}\\ &= (MmD\Phi_1^2 + M^2D\Phi_1\Phi_2 + m^2d\Phi_1\Phi_2 + Mmd\Phi_2^2)\frac{1}{\lambda_p},\end{aligned} \tag{3.24}$$

$$\begin{aligned}(mD\Phi_1 + Md\Phi_2)p_j &= (mD\Phi_1 + Md\Phi_2)(M\Phi_1 + m\Phi_2)\frac{1}{\lambda_p}\\ &= (MmD\Phi_1^2 + m^2D\Phi_1\Phi_2 + M^2d\Phi_1\Phi_2 + Mmd\Phi_2^2)\frac{1}{\lambda_p},\end{aligned} \tag{3.25}$$

$$\begin{aligned}p_jp_k &= (M\Phi_1 + m\Phi_2)(m\Phi_1 + M\Phi_2)\frac{1}{\lambda_p^2}\\ &= (Mm\Phi_1^2 + M^2\Phi_1\Phi_2 + m^2\Phi_1\Phi_2 + Mm\Phi_2^2)\frac{1}{\lambda_p^2}.\end{aligned} \tag{3.26}$$

Substituting (3.24)–(3.26) into (3.23) we have

$$\frac{1}{\lambda_p}(M^2D\Phi_1\Phi_2 + m^2d\Phi_1\Phi_2 - m^2D\Phi_1\Phi_2 - M^2d\Phi_1\Phi_2)$$
$$= \frac{\lambda}{\lambda_p^2}(Mm\Phi_1^2 + M^2\Phi_1\Phi_2 + m^2\Phi_1\Phi_2 + Mm\Phi_2^2)(D-d),$$

$$M^2\Phi_1\Phi_2(D-d) - m^2\Phi_1\Phi_2(D-d) = \frac{\lambda}{\lambda_p}(Mm\Phi_1^2 + M^2\Phi_1\Phi_2 + m^2\Phi_1\Phi_2 + Mm\Phi_2^2)(D-d),$$

$$\frac{M+m}{M-m}(M^2-m^2)\Phi_1\Phi_2 = Mm\Phi_1^2 + M^2\Phi_1\Phi_2 + m^2\Phi_1\Phi_2 + Mm\Phi_2^2,$$

$$(M^2 + 2Mm + m^2)\Phi_1\Phi_2 = Mm\Phi_1^2 + M^2\Phi_1\Phi_2 + m^2\Phi_1\Phi_2 + Mm\Phi_2^2,$$

$$2Mm\Phi_1\Phi_2 = Mm\Phi_1^2 + Mm\Phi_2^2,$$

$$\Phi_1^2 - 2\Phi_1\Phi_2 + \Phi_2^2 = 0,$$

$$(\Phi_1 - \Phi_2)^2 = 0.$$

Hence

$$\Phi_1 = \Phi_2.$$

LEMMA 3.13. *Let A be an $n \times n$ positive matrix with real eigenvalue λ satisfying* (3.1) *and let f be normalized with respect to j and k. If $\lambda > 0$, then*

$$\nu f_l = \begin{cases} p_l & \text{if } l \in \mathscr{J}, \\ -p_l & \text{if } l \in \mathscr{K} \end{cases}$$

and if $\lambda < 0$, then

$$\nu f_l = \begin{cases} -p_l & \text{if } l \in \mathscr{J}, \\ p_l & \text{if } l \in \mathscr{K} \end{cases}$$

for some positive number ν.

Proof. Suppose $\lambda > 0$; the proof in the other case is similar. The matrix equations $(Af)_j = \lambda f_j$ and $(Ap)_j = \lambda_p p_j$ are equivalent to the following:

$$\sum_{l \in \mathscr{J}} a_{jl} f_l + \sum_{l \in \mathscr{K}} a_{jl} f_l = \lambda f_j \tag{3.27}$$

and

$$\sum_{l \in \mathscr{J}} a_{jl} p_l + \sum_{l \in \mathscr{K}} a_{jl} p_l = \lambda_p p_j. \tag{3.28}$$

Using the definitions of Φ_1, Φ_2, $\mathscr{J}$ and $\mathscr{K}$, (3.27) and (3.28) become, respectively,

$$MD\Phi_1 + md\Phi_2 = \lambda D p_j,$$

$$M\Phi_1 + m\Phi_2 = \lambda_p p_j.$$

Then

$$\Phi_1 = \frac{p_j(\lambda D - \lambda_p d)}{M(D-d)}$$

and

$$\Phi_2 = \frac{p_j D(\lambda_p - \lambda)}{m(D-d)}.$$

But by Lemma 3.12, $\Phi_1 = \Phi_2$ and thus

$$\frac{p_j(\lambda D - \lambda_p d)}{M(D-d)} = \frac{p_j D(\lambda_p - \lambda)}{m(D-d)},$$

$$m(\lambda D - \lambda_p d) = MD(\lambda_p - \lambda),$$

$$\lambda Dm - \lambda_p md = \lambda_p MD - \lambda MD,$$

$$\lambda Dm + \lambda MD = \lambda_p MD + \lambda_p md,$$

$$\lambda D(m+M) = \lambda_p(MD + md),$$

$$\frac{M-m}{M+m}\lambda_p D(m+M) = \lambda_p D(M-m) = \lambda_p(MD + md),$$

$$D(M-m) = MD + md,$$

$$-Dm = md,$$

$$D = -d.$$

THEOREM 3.1. *Let A be an $n \times n$ positive matrix with eigenvalue λ satisfying* (3.1) *with $M > m$. Then n is even,*

$$\lambda_p = \frac{n}{2}(M+m), \tag{3.29a}$$

$$\lambda = \pm\frac{n}{2}(M-m), \tag{3.29b}$$

and $(1/\lambda_p)A$ *is a stochastic matrix with the following structure.*

(a) *If* $\lambda > 0$, *there exists a permutation matrix P so that*

$$PAP^T = \begin{bmatrix} \bar{M} & \bar{m} \\ \bar{m} & \bar{M} \end{bmatrix} \tag{3.30}$$

where

$$\bar{M} = \begin{bmatrix} M & \cdots & M \\ \vdots & & \vdots \\ M & \cdots & M \end{bmatrix} \tag{3.31}$$

and

$$\bar{m} = \begin{bmatrix} m & \cdots & m \\ \vdots & & \vdots \\ m & \cdots & m \end{bmatrix} \tag{3.32}$$

are square matrices of order $n/2$. The normalized eigenvector Pf corresponding to λ has the form

$$Pf = \begin{cases} 1 & \text{if } 1 \leqq i \leqq \dfrac{n}{2}, \\ -1 & \text{if } \dfrac{n}{2} \leqq i \leqq n. \end{cases} \tag{3.33}$$

(b) *If* $\lambda < 0$, *there exists a permutation matrix P so that*

$$PAP^T = \begin{bmatrix} \bar{m} & \bar{M} \\ \bar{M} & \bar{m} \end{bmatrix}$$

where $\bar{m}$ and $\bar{M}$ are defined in (3.31) *and* (3.32). *The normalized eigenvector Pf corresponding to λ has the form*

$$Pf = \begin{cases} -1 & \text{if } 1 \leqq i \leqq \dfrac{n}{2}, \\ 1 & \text{if } \dfrac{n}{2} \leqq i \leqq n. \end{cases}$$

Proof. By Lemma 3.9 the eigenvalue λ is real. Suppose $\lambda > 0$; the proof for part (b) is similar. Let the sets $\mathcal{J}$ and $\mathcal{K}$ be given by $\{j_1, \cdots, j_\gamma\}$ and $\{k_1, \cdots, k_{n-\gamma}\}$, respectively, and let P be the permutation matrix corresponding to the permutation $\{j_1, \cdots, j_\gamma, k_1, \cdots, k_{n-\gamma}\}$. By Lemma 3.11, PAP^T must have the structure in (3.30)–(3.32). The equations $(Ap)_j = \lambda_p p_j$ and $(Ap)_k = \lambda_p p_k$ with $j \in \mathcal{J}$ and $k \in \mathcal{K}$ are equivalent to

$$M\Phi_1 + m\Phi_2 = \lambda_p p_j \tag{3.34}$$

and

$$m\Phi_1 + M\Phi_2 = \lambda_p p_k. \tag{3.35}$$

By Lemma 3.11 we may sum (3.34) and (3.35) over indices j, k to obtain

$$\gamma M\Phi_1 + \gamma m\Phi_2 = \lambda_p \Phi_1 \tag{3.36}$$

and

$$(n-\gamma)m\Phi_1 + (n-\gamma)M\Phi_2 = \lambda_p \Phi_2. \tag{3.37}$$

Applying Lemma 3.12 to (3.36) and (3.37) we have

$$\gamma M + \gamma m = \lambda_p, \qquad (n-\gamma)m + (n-\gamma)M = \lambda_p.$$

Then

$$\gamma = \frac{\lambda_p}{M+m} \tag{3.38}$$

and

$$n-\gamma = \frac{\lambda_p}{M+m}.$$

Therefore,

$$\gamma = n - \gamma = \frac{n}{2}. \tag{3.39}$$

The expressions for λ_p and λ in (3.29) follow immediately from (3.38) and (3.39). Using (3.34), (3.35) and Lemma 3.12, we may write $p_j = p_k$ and p as

$$p = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Thus (3.33) is true.

REFERENCES

[1] E. Hopf, *An inequality for positive integral operators*, J. Math. Mech., 12 (1963), pp. 683–692.

[2] A. M. Ostrowski, *Positive matrices and functional analysis*, in Recent Advances in Matrix Theory, H. Schneider, ed., University of Wisconsin Press, Madison–Milwaukee, WI, 1964.

[3] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[4] U. C. Rothblum and C. P. Tan, *Upper bounds on the maximum modulus of subdominant eigenvalues of nonnegative matrices*, Linear Algebra Appl., 66 (1985), pp. 45–86.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

017

THE GENERAL MINIMUM FILL-IN PROBLEM*

H. WENDEL†

Abstract. We consider the well-known graph-theoretical elimination process which is related to Gaussian elimination on a sparse, positive definite system of linear equations. The general minimum fill-in problem is concerned in ranking (elimination) orderings by so-called criterion functions and is interested in those orderings which minimize (which are optimal with respect to) any (fixed) criterion function or, more generally, which minimize even the whole class of such functions. A valuable tool for attacking this problem is the Initial Theorem due to Bertele and Brioschi (J. Math. Anal. Appl., 35 (1971), pp. 48–57). An Isomorphic Theorem can be proved guaranteeing a particular invariance property which is of great importance for the application of the Initial Theorem. In addition we consider the so-called separation approach which—roughly speaking—splits a given graph G into two partial graphs G_1 and G_2 so that optimal orderings of G_1 and G_2 together form an optimal ordering of G. We are able to give conditions on a separating set of vertices sufficient for this procedure. Furthermore, a special class of graphs is introduced which arise in the field of load-flow calculation. The Initial Theorem is generalized to that class of graphs.

Key words. minimum fill-in problem, optimal ordering, Gaussian elimination, sparse linear equations, load-flow calculation, separating set, tearing a graph

1. Introduction and notation. The graph-theoretic minimum fill-in problem is of great interest in the field of sparse matrix research. It is stated as follows: Given an undirected and finite graph $G = (X, E)$, $|X| =: n$, without self-loops and parallel edges (only those graphs are considered in this paper), X is the set of vertices and E is a set of unordered pairs (x, y) of distinct vertices, the set of edges. For G the well-known elimination of a vertex x is defined by

— adding edges to the neighbourhood Adj $(x|G)$ of x (Adj $(x|G)$ denotes the set of vertices adjacent to x) so that Adj $(x|G)$ becomes a clique;

— deleting x along with the edges belonging to x from G.

The graph resulting from this elimination process is denoted by $G_x = (X_x, E_x)$. Every ordering α of X (which should be considered as a sequence $\alpha = \langle \alpha(1), \cdots, \alpha(n) \rangle$ of distinct vertices of G, respectively, as a bijective map α: $\{1, \cdots, n\} \rightarrow X$) induces a sequence of elimination graphs

$$G_{\langle \alpha(1), \cdots, \alpha(k) \rangle} := (G_{\langle \alpha(1), \cdots, \alpha(k-1) \rangle})_{\alpha(k)}, \qquad k=0, \cdots, n.$$

In this context sequences of distinct vertices are called *elimination orderings*. The *fill-in* $F(\alpha)$ produced by an elimination ordering α is the set of new edges which arise during the successive elimination process. The problem of determining elimination orderings α producing a minimal fill-in $F(\alpha)$ (in the sense of $|F(\alpha)| \leqq |F(\alpha')|$ for all elimination orderings α' of G) is well known as the minimum fill-in problem.

In this paper a generalization of this problem, introduced by Rose [Ro70], is considered. The so-called general minimum fill-in problem deals with ranking elimination orderings by *criterion functions*. By definition, a criterion function is a function f: $\mathbb{N}^n \rightarrow \mathbb{R}$ which satisfies

$$f(a_1, \cdots, a_n) = f(a_{\pi(1)}, \cdots, a_{\pi(n)}) \quad \text{for all permutations } \pi \text{ of the numbers } 1, \cdots, n$$

and

$$a_i \leqq b_i \quad \text{for } 1 \leqq i \leqq n \Rightarrow f(a_1, \cdots, a_n) \leqq f(b_1, \cdots, b_n).$$

* Received by the editors August 14, 1985; accepted for publication May 18, 1987.

† Institut für Informatik, Technische Universität München, Arcisstraße 21, D-8000 München 2, West Germany.

The class of criterion functions, suitable for G, is denoted by $\mathscr{F}_n$, respectively, $\mathscr{F}(G)$. If there is no doubt about the underlying graph G we write $\mathscr{F}$ only. In order to relate criterion functions to an elimination ordering α, we attach to α the vector

$$D(\alpha|G) := (d(\alpha(i)|G_{\langle\alpha(1), \cdots, \alpha(i-1)\rangle}))_{i=1,\cdots,n}$$

(where $d(x|G)$ denotes the degree of the vertex x in the graph G). The entries of $D(\alpha|G)$ are the degrees of the vertices $\alpha(i)$ (at the moment of their elimination) in the elimination graph $G_{\langle\alpha(1), \cdots, \alpha(i-1)\rangle}$. Note that $D(\alpha|G)$ can be regarded as the "result" of α. Ranking an elimination ordering by a criterion function f is carried out by calculating $f(D(\alpha|G))$.

The general minimum fill-in problem is stated by the following two types of optimality.

DEFINITION 1.1. An elimination ordering α of G is called *f-minimal* ($f \in \mathscr{F}$), if $f(D(\alpha|G)) \leqq f(D(\alpha'|G))$ for all elimination orderings α' of G. We call α *$\mathscr{F}$-minimal*, if α is f-minimal for all $f \in \mathscr{F}$.

Evidently, f-minimal elimination orderings always exist. In contrast, the existence of $\mathscr{F}$-minimal elimination orderings cannot be guaranteed. But note that in [We83] such elimination orderings are determined for (large) graphs which arise in connection with load-flow calculation in power systems.

Two criterion functions are of particular interest in the field of sparse matrix computation. These are

$$f_L(a_1, \cdots, a_n) := \sum_{i=1}^{n} a_i \quad \text{and} \quad f_Q(a_1, \cdots, a_n) := \sum_{i=1}^{n} a_i^2.$$

The function f_L can be used to give an alternative definition of the minimum fill-in problem: An elimination ordering α minimizes the fill-in $F(\alpha)$ if and only if α is f_L-minimal. This equivalence justifies the name "general minimum fill-in problem."

The elimination process on graphs models the combinatorial features of Gauss' (respectively, Cholesky's) algorithm for solving a sparse and positive definite system of linear algebraic equations $Mx = b$; a detailed analysis of this subject is given in [Ro70] and [Ro72]. For example, the factorization of the matrix M requires $f_1(D(\alpha|G))$ multiplications and $f_2(D(\alpha|G))$ additions, where $G = G(M)$ is the graph representing the zero-nonzero pattern of M, α is an elimination ordering of G which corresponds to the used sequence of diagonal pivots and f_1 and f_2 are criterion functions defined by $f_1 := 0.5 \cdot f_Q + 1.5 \cdot f_L$, $f_2 := 0.5 \cdot f_Q + 0.5 \cdot f_L$. Analogously, the number of multiplications and the number of additions for the backsolving process are given by $2 \cdot f_L(D(\alpha|G))$, the space to store the triangular factors of M by $f_L(D(\alpha|G)) + n$. Summarizing, we see that an f-minimal elimination ordering α of $G = G(M)$, where f is one of the criterion functions defined above, corresponds to a sequence of diagonal pivots which minimizes the corresponding arithmetic, respectively, space, criterion. More generally an $\mathscr{F}$-minimal elimination ordering corresponds to an ordering of M which is optimal with respect to both, arithmetic and space requirements.

In addition to criterion functions, Rose [Ro70], [Ro72] and Bertelè and Brioschi [Be71], [Be72] have introduced a quasi-ordering relation for ranking elimination orderings.

DEFINITION 1.2. For two elimination orderings α and α' of a graph

$$G = (X, E), n := |X|,$$

it is said that α *dominates* α' if there is a permutation π of the numbers $1, \cdots, n$ with

$$d(\alpha(i)|G_{\alpha(1,i-1)}) \leqq d(\alpha'(\pi(i))|G_{\alpha'(1,\pi(i)-1)}) \quad \text{for } i = 1, \cdots, n.$$

α is called *dominating* if it dominates all elimination orderings of G. If there exists a permutation π of the numbers $1, \cdots, n$ with

$$d(\alpha(i)|G_{\alpha(1,i-1)}) = d(\alpha'(\pi(i))|G_{\alpha'(1,\pi(i)-1)}) \quad \text{for } i = 1, \cdots, n$$

we say that α and α' are *equivalent* to each other.

Dominating orderings exist for chordal graphs [Ro70]. But there are graphs with no such ordering [Be72, Example 2.7.2]. In Appendix A2 it is proved that an elimination ordering α is $\mathscr{F}$-minimal if and only if it is dominating. In particular, we will see that α is $\mathscr{F}$-minimal if and only if it is f-minimal for only a finite number of suitable criterion functions f.

It is well known that there is no general and feasible procedure for solving the minimum fill-in problem. The reason for this lack is provided by Yannakakis [Ya81] which proves this problem to be NP-complete. Therefore, any attempt to solve the general minimum fill-in problem may only result in methods which do not succeed in the whole variety of graphs. But we note that from the theoretical point of view we are interested in any procedure to determine f-minimal elimination orderings. This is because, in the field of sparse matrix computation, standards are required for finding out the absolute quality of the heuristical ordering procedures which are currently in use (for example, the minimum degree algorithm, the minimum deficiency algorithm, the reverse Cuthill-MacKee algorithm, the nested dissection algorithm and the one-way dissection algorithm), i.e., we want to know how close the heuristic orderings are to the theoretical figures of the minimum fill-in and the minimal number of arithmetic operations. In [We83], for some large graphs out of the field of load-flow calculation in electrical power systems, those theoretical figures are computed and listed along with the corresponding results of the heuristic procedures. As an interesting result of this classification we mention that the minimum degree algorithm really produces "near-optimal" orderings (at least in the considered field of application). This paper presents the methods used to compute these orderings.

The most useful and efficient approach is (in the author's opinion) that of Bertelè and Brioschi [Be71], [Be72]. Its key idea is the so-called Initial Theorem: If G contains a (so-called) vertex x of type B then there exists an f-minimal elimination ordering (for any $f \in \mathscr{F}$) of G starting with x. Repeated application of the Initial Theorem (if possible) leads either to a $\mathscr{F}$-minimal elimination ordering or at least to a starting sequence of an f-minimal elimination ordering. In the second case an isomorphic theorem (Theorem 2.6) is proved; it guarantees that all those starting sequences are "equivalent." Consequently, the repetitive elimination process can be realized in a very efficient algorithmic manner. It is obvious that a graph need not have any vertex of type B and therefore the approach of Bertelè and Brioschi may fail. But this situation confirms the nonexistence of a general and feasible method to solve the general minimum fill-in problem, rather than being a disadvantage of the method itself.

Another method used to attack our problem is the so-called separation approach, which is motivated by a well-known principle in optimization theory stating that a large optimization problem should be solved (if possible) by separating it into smaller subproblems. In connection with the determination of optimal sequences of pivots such separation procedures (among others) are given in [Bau67], [Ti73], [Ge73], [Ge78], [Ge80]. Common to all these methods is that a given graph G is split into partial graphs $G_i = G(X_i)$ by removing a separating set S of vertices. For the partial graphs G_i, near-optimal (in general only "good") elimination orderings β_i are constructed. Subsequently, these orderings are concatenated to an elimination ordering $\alpha = \beta_1 + \cdots + \beta_n + \delta$ of G, where δ is any suitable ordering of S (+ denotes the concatenation operator). Commonly, the subdivision of the graph is motivated by a great deal of heuristics (for example, the subdivision of a power system in local subsystems): global optimality is out of consideration. Therefore none of these methods cited above can guarantee that the computed results minimize the considered criterion exactly; yet some of them produce results which are optimal in an asymptotic sense.

In contrast we are only interested in those subdivisions of $G = (X, E)$ for which the elimination orderings β_i and δ form a f-minimal elimination ordering of G. The basic ideas of this separation approach and its application are introduced in § 3. Obviously, splitting a graph with respect to f-minimality imposes hard restrictions on the related separating set of vertices. In § 4, three separation theorems are presented which introduce such conditions. The proof of one of these theorems is given in detail in the Appendix. It should be noted that the method of Bertelè and Brioschi can be considered as a special case of separation, respectively, the Initial Theorem is a straightforward corollary of one of the mentioned separation theorems.

The idea of separation to solve the general minimum fill-in problem has already been proposed by Rose [Ro70] and Bertelè and Brioschi [Be71]. Their ideas are summarized by the following three theorems which are written down in our notation.

THEOREM 1.3. *Let $G = (X, E)$ be a graph and $S \subset X$ be a complete set of vertices of G splitting G into the two (partial) graphs $G(Y)$ and $G(Z)$, where $X \backslash S = Y \dot{\cup} Z$ (i.e., $G(X \backslash S) = G(Y) \oplus G(Z)$). Then to each $f \in \mathscr{F}$ there exists an f-minimal elimination ordering α of G (depending on f) eliminating first the vertices of Y followed by the vertices of Z; the vertices of S are eliminated at the end.*

Theorem 1.3 follows directly from Theorems 1.4 and 1.5, which are frequently used in this paper.

THEOREM 1.4 (Final Theorem). *Let $G = (X, E)$ be a graph and $S \subset X$ be a complete set of vertices of G. Then for each $f \in \mathscr{F}$ there exists an f-minimal elimination ordering α of G (depending on f) eliminating the vertices of S at the end, i.e., $\alpha(i) \in S$ for $i = |X| - |S| + 1, \cdots, |X|$.*

THEOREM 1.5. *Let $G = (X, E)$ be a graph and $S \subset X$ be a separating set of vertices splitting G into the two (partial) graphs $G(Y)$ and $G(Z)$, where $X \backslash S = Y \dot{\cup} Z$ (i.e., $G(X \backslash S) = G(Y) \oplus G(Z)$). Furthermore let α and α' be two elimination orderings satisfying*

— $\alpha(i), \alpha'(i) \in S$ for $i = |X| - |S| + 1, \cdots, |X|$ (i.e., S is eliminated at the end);

— $\alpha_Y = \alpha'_Y$ and $\alpha_Z = \alpha'_Z$.

Then α and α' are equivalent to each other.

In the field of load-flow calculation in power systems a special type of graphs, called simplex graphs, is of interest. In § 5 we will consider f, respectively, $\mathscr{F}$, minimal elimination orderings of such graphs. The Initial Theorem is generalized to simplex graphs. Finally, Appendix A1 summarizes some statements and "rules of computation" for handling elimination graphs.

The rest of this section introduces some additional notation. For an elimination ordering $\alpha = \langle \alpha(1), \cdots, \alpha(n) \rangle$ of a graph $G = (X, E)$, $|X| = n$, a partial sequence $\langle \alpha(i_1), \cdots, \alpha(i_k) \rangle =: \beta = \langle \beta(1), \cdots, \beta(k) \rangle$ of α, where $1 \leqq i_1 < i_2 < \cdots < i_k \leqq n$ is called *partial elimination ordering*. The length k of β is denoted by $|\beta|$, the set of vertices belonging to β by $M(\beta) := \{\alpha(i_1), \cdots, \alpha(i_k)\}$. For the empty elimination ordering we write $\langle \quad \rangle$. Special partial elimination orderings of α are the so-called "sections" of α, $\alpha(i, j) := \langle \alpha(i), \alpha(i + 1), \cdots, \alpha(j) \rangle$, $i, j \in \mathbb{N}$. Evidently, $\alpha(i, j)$ is empty for $i > j$, respectively, for $i > |\alpha|$. For a subset $S \subset X$, α_S denotes the "S-part" of α, i.e., that partial elimination ordering of α consisting of all those $\alpha(j)$ with $\alpha(j) \in S$. In other words, α_S arises from α by erasing all $\alpha(j)$ with $\alpha(j) \notin S$ from α. The concatenation $\beta + \gamma$ of two partial elimination orderings β, γ with $\ell := |\beta|$ and $m := |\gamma|$ is defined by

$$\beta + \gamma := \langle \beta(1), \cdots, \beta(\ell), \gamma(1), \cdots, \gamma(m) \rangle,$$

their "difference" by $\beta - \gamma := \beta_{M(\beta)\setminus M(\gamma)}$. If $M(\beta) \cap M(\gamma) = \emptyset$, $\beta + \gamma$ becomes a (partial) elimination ordering.

Now we introduce further graph-theoretical notation. In addition to Adj $(x|G)$ the neighbourhood of a subset $Y \subset X$ is defined by

$$\mathrm{Adj}\,(Y|G) := \{x \in X \setminus Y | (x, y) \in E \text{ for any } y \in Y\}.$$

For $Y \subset X$ the *section graph with respect to* Y is denoted by $G(Y)$. If there is no doubt we will relate graph-theoretical notions directly to the set Y instead of onto $G(Y)$ (for example: Y complete instead of $G(Y)$ complete). The *Y-elimination graph* of G is denoted by $G_Y = (X_Y, E_Y)$ and (well-) defined by eliminating the vertices of Y in any order. A sequence $\omega = \langle x_0, \cdots, x_\ell \rangle$ of distinct vertices of $G = (X, E)$ satisfying $(x_i, x_{i+1}) \in E$ for $0 \leqq i \leqq \ell - 1$ is called a *path* of length ℓ from x_0 to x_ℓ. If $\ell = 1$ we speak of a trivial path. For the set of intermediate vertices of a path ω we introduce the notation $Z(\omega) := \{x_1, \cdots, x_{\ell-1}\}$. A path ω_1 from a_1 to b_1 is called *disjunct* to a path ω_2 from a_2 to b_2, if $Z(\omega_1) \cap Z(\omega_2) = \emptyset$ and $a_1, b_1 \notin Z(\omega_2)$ and $a_2, b_2 \notin Z(\omega_1)$. A system of paths ω_j from a_j to b_j, $j = 1, \cdots, k$, is called disjunct if the paths are mutual disjunct. For two graphs $G_1 = (X_1, E_1)$, $G_2 = (X_2, E_2)$ with $X_1 \cap X_2 = \emptyset$ we define the *direct sum* of G_1 and G_2 by $G_1 \oplus G_2 := (X_1 \dot{\cup} X_2, E_1 \dot{\cup} E_2)$, where $\dot{\cup}$ denotes the union of disjunct sets. By this notion, for any separating set $S \subset X$ there is a partition $Y \dot{\cup} Z \dot{\cup} S = X$ of the set of vertices of $G = (X, E)$ so that $G(X \setminus S) = G(Y) \oplus G(Z)$. The complementary graph of a graph G is denoted by $\bar{G}$. Furthermore the terms "complete graph" and "clique" are used synonymously.

2. The approach of Bertelè and Brioschi. The basic idea of the approach of Bertelè and Brioschi is the so-called Initial Theorem. The assumptions of the Initial Theorem are stated by the following definition. But first we remember that a graph $G = (X, E)$ is called a *bush* if there is a vertex w satisfying: $d(w|G) = |X| - 1$ and $d(b|G) = 1$ for all $b \in X \setminus \{w\}$; the vertex w is called *root*, the vertices $b \in X \setminus \{w\}$ are denoted as *peaks*. If $|X| = 1$, we speak of a trivial bush, consisting of a root only. A graph G is called a *forest of bushes* if it is the direct sum of bushes.

DEFINITION 2.1. Let $G = (X, E)$ be a graph. A vertex x is called of type B if:

(i) $\overline{G(\mathrm{Adj}\,(x|G))}$ is a forest of bushes.

(ii) To each of these bushes with root w and peaks $b_1, \cdots, b_k$ there are disjunct paths ω_i, $i = 1, \cdots, k$ in G from w to b_i with $Z(\omega_i) \cap (\mathrm{Adj}\,(x|G) \cup \{x\}) = \emptyset$.

Figure 1 shows some typical samples of vertices x of type B together with their neighbourhood. Paths are indicated by -----. Evidently, any vertex having a complete neighbourhood is of type B.

THEOREM 2.2 (Initial Theorem [Be72, Thm. 3.5.2]). *Let $G = (X, E)$ be a graph and $x \in X$ be a vertex of type* B. *Then to each $f \in \mathcal{F}$ there exists an f-minimal elimination ordering α of G (depending on f) starting with x, i.e., with $\alpha(1) = x$.*

Analogously to the Final Theorem the Initial Theorem follows directly from Theorem 2.3.

THEOREM 2.3 [Be72, Thm. 3.5.1]. *Let $G = (X, E)$ be a graph and $x \in X$ be a vertex of type* B. *Then to each elimination ordering α' of G there exists an elimination ordering α of G which dominates α' and starts with x.*

Repetitive elimination of vertices of type B is described by the following.

DEFINITION 2.4. A (partial) elimination ordering α of a graph G is called a B-*(partial) elimination ordering*, if $\alpha(i)$ is in $G_{\alpha(1,i-1)}$ of type B, $1 \leqq i \leqq |\alpha|$. We call α *noncontinuable* if G_α does not contain any vertex of type B.

B-elimination orderings exist for chordal graphs. This is verified by induction over the number of vertices, where we have to apply Dirac's Lemma, which guarantees that

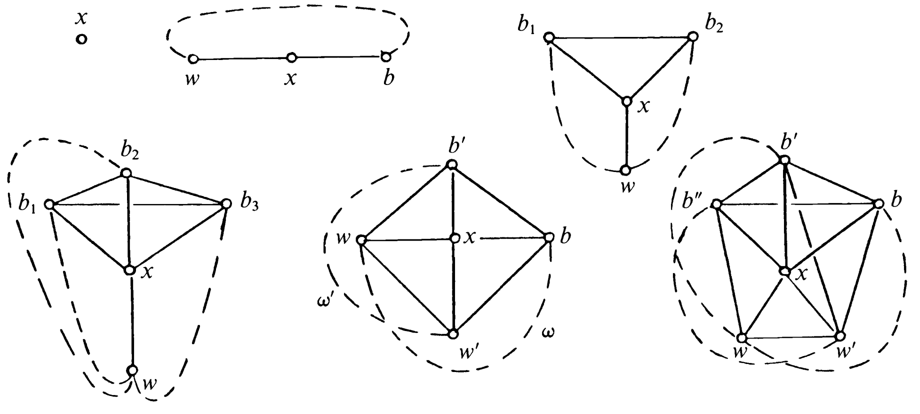


FIG. 1. *Note that ω and ω' may not be disjunct.*

a chordal graph always contains a vertex with a complete neighbourhood. The key property of B-(partial) elimination orderings is described by the following proposition.

PROPOSITION 2.5. (i) *Every* B-*elimination ordering of a graph is $\mathscr{F}$-minimal.*

(ii) *For any $f \in \mathscr{F}$ it is true that: A* B-*partial elimination ordering β of a graph G can be chosen as a starting sequence of a f-minimal elimination ordering,* i.e., *there exists a partial elimination ordering γ (depending on f) so that $\beta + \gamma$ is a f-minimal elimination ordering of G.*

Since $\mathscr{F}$-minimal elimination orderings may not exist, a B-partial elimination ordering of a graph is not necessarily a starting sequence of a $\mathscr{F}$-minimal elimination ordering.

Example. Consider the graph G given in Fig. 2 ([Ro70]). Then

$$\alpha := \langle 1, 2, 3, \cdots, 11 \rangle$$

is a B-elimination ordering with $f_L(D(\alpha|G)) = 27$ and $f_Q(D(\alpha|G)) = 77$. Other graphs for which a B-elimination ordering exists are presented in [We83]. Most of them are of interest in the field of load-flow calculation in power systems.

Now we consider the situation that for a graph G the successive elimination of vertices of type B leads to a (real) noncontinuable B-partial elimination ordering. This situation arises when, during the process of eliminating vertices of type B, an elimination graph has been generated that does not contain any more type-B vertices. From the theoretical as well as from the practical point of view, it is interesting to know whether any other (larger) B-partial elimination ordering exists. If such an ordering existed (and this is absolutely imaginable) the determination of the "best" B-partial elimination or-

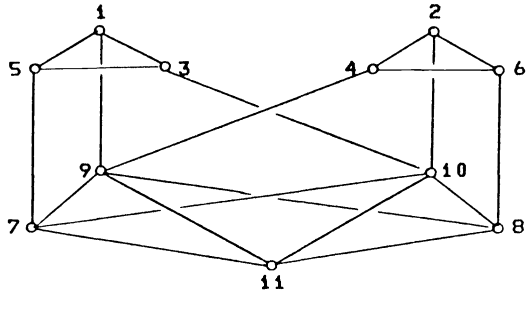


FIG. 2

dering would require a backtracking process in order to inspect all noncontinuable B-partial elimination orderings. But fortunately the length of noncontinuable B-partial elimination orderings is an invariable of a graph G. This is guaranteed by the following theorem.

THEOREM 2.6. *For any two noncontinuable* B-*partial elimination orderings* β *and* β' *of a graph* G *it is true that*

(i) $|\beta| = |\beta'|$;

(ii) G_β *is isomorphic to* $G_{\beta'}$.

The proof of Theorem 2.6 is given in Appendix A3.

Obviously, a graph G has no B-elimination ordering if and only if there exists at least one true B-partial elimination ordering of G which is noncontinuable. In particular, all noncontinuable B-partial elimination orderings of G are elimination orderings if there exists a B-elimination ordering of G. For example any repetitive elimination process of type-B-vertices in the graph of Fig. 2 always terminates in the empty graph. Yet the most important consequence of Theorem 2.6 is that "the best" B-partial elimination ordering can be determined by a simple and backtracking-free algorithm which is sketched below.

```
$\alpha := \langle\ \rangle$;
$G^{(0)} := G$;
$i := 0$;
WHILE '$G^{(i)}$ contains a vertex of type B' DO
BEGIN
choose any vertex $x$ of type B out of $G^{(i)}$;
$G^{(i+1)} := (G^{(i)})_x$;
$\alpha := \alpha + \langle x\rangle$;
$i := i + 1$;
END
```

In implementing this algorithm a test-procedure is required to check whether a vertex is of type B or not. In this procedure the verification of condition 2.1 (i) is easily done. In contrast the test of condition 2.1 (ii) is a much more difficult task. In order to carry out this verification in an efficient manner it should be realized as a (equivalent) flow problem. For solving the flow problem efficient procedures are available.

The author has used a PASCAL-implementation of this algorithm; the involved flow problem is solved by the algorithm of Dinic [Ev75]. The program works very efficiently. For example the computation of a B-elimination ordering for a graph G containing 118 vertices (for readers familiar with power systems, it is the AEP 118-bus test network) requires 4.9 seconds' execution time on a DEC PDP-11/23.

3. The separation approach: basic ideas. We begin this section with an example introducing some problems concerning any separation approach. Let us look at the graph $G = (X, E)$ of Fig. 3 for which a f_L-minimal elimination ordering should be determined ($f_L(a_1, \cdots, a_n) := \sum a_i$). We see that G does not contain any vertex of type B and therefore the method of Bertelè and Brioschi fails. According to Theorem 1.3 an f_L-minimal elimination ordering α of G exists eliminating first the vertices $Y = \{y_1, \cdots, y_4\}$ succeeded by the vertices $Z = \{z_1, \cdots, z_4\}$; the vertices $S = \{s_1, \cdots, s_4\}$ are eliminated at the end of α. In order to determine α in detail (Theorem 1.3 guarantees the existence of α only) we compute, for example by a combinatorial procedure [Bau67], a partial elimination ordering β of G which satisfies $M(\beta) = Y$ and which minimizes

$$\sum_{i=1}^{4} d(\beta(i)|G_{\beta(1,i-1)})$$

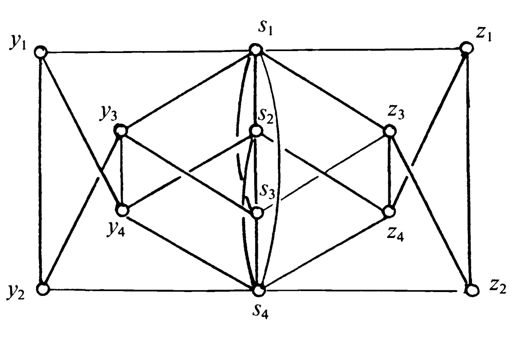


FIG. 3

(roughly speaking, an f_L-minimal partial elimination ordering β with $M(\beta) = Y$). Subsequently, also by a combinatorical process, an f_L-minimal elimination ordering γ of G_Y is computed which eliminates the vertices of S at the end. Evidently, the elimination ordering $\beta + \gamma$ is f_L-minimal. The "algorithmic expense" has been reduced by this procedure from $|X|!$ to $|Y|! + |Z|! + |S|!$, respectively, from $2^{|X|}$ to $2^{|Y|} + 2^{|Z|} + 2^{|S|}$.

Problems arise in the notation when we have to minimize an arbitrary and more complicated criterion function f; a good example of such a function is given in Appendix A2. Again, Theorem 1.3 guarantees the existence of an f-minimal elimination ordering α eliminating Y, Z and S one by one. But in this situation we are not able to construct an f-minimal elimination ordering by minimizing f in the partial manner as it has been done above because we do not know the "kind of optimality" which matches the constituents β and γ of α. An imaginable but expensive alternative would be to compute $f(D(\alpha|G))$ for all elimination orderings α composed of β and γ as mentioned above and to choose an f-minimal one; this procedure reduces the algorithmic "expense" from $|X|!$ to $|Y|! \cdot |Z|! \cdot |S|!$ only.

Now a notation is introduced appropriate for handling subdivisions of graphs in connection with f, respectively, $\mathcal{F}$, minimal elimination orderings.

DEFINITION 3.1. Let $G = (X, E)$ be a graph and $f \in \mathcal{F}$. G is called *separable with respect to f* if there is a separating set $S \subset X$ splitting G into two partial graphs $G(Y)$ and $G(Z)$, where $X \backslash S = Y \dot{\cup} Z$ (i.e., $G(X \backslash S) = G(Y) \oplus G(Z)$). Furthermore there exists a corresponding f-minimal elimination ordering α of G satisfying

$$\alpha(i) \in \begin{cases} Y & \text{for } i = 1, \cdots, |Y|, \\ Z & \text{for } i = |Y|+1, \cdots, |Y|+|Z|, \\ S & \text{for } i = |Y|+|Z|+1, \cdots, |X|. \end{cases}$$

The partition (Y, S, Z) of X is called *decomposition of G with respect to f.* A partition (Y, S, Z) which is a decomposition of G with respect to all $f \in \mathcal{F}$ is called a *decomposition of G with respect to $\mathcal{F}$.*

In § 4 sufficient conditions on a graph G are established guaranteeing that G is separable with respect to $\mathcal{F}$.

Now the lack of our notation which has appeared in the introductory example is eliminated. Essentially, we have to specify which "kind of optimality" is satisfied by a partial elimination ordering of an f-minimal elimination ordering. It will become evident that ranking partial elimination orderings by criterion functions is appropriate too. Therefore, we define the (well-known) vector

$$D(\beta|G) := (d(\beta(i)|G_{\beta(1,i-1)}))_{i=1,\cdots,\ell}$$

also for partial elimination orderings $\beta = \langle\beta(1), \cdots, \beta(\ell)\rangle$ of a graph G. Comparing two partial elimination orderings β and β' of a graph G is reasonable only if $M(\beta) =$

$M(\beta')$; for example, the part β of an elimination ordering $\alpha = \beta + \gamma$ can be replaced by a possibly "better" partial elimination ordering β' only if $M(\beta) = M(\beta')$. In this context for a partial elimination ordering β of G satisfying $M(\beta) = Y$, $Y \subset X$, we introduce the notation *Y-elimination ordering of G*. For Y-elimination orderings the terms f, respectively, $\mathscr{F}$ minimal are used analogously, i.e., a Y-elimination ordering β is called *f-minimal* ($f \in \mathscr{F}_\ell$, $\ell = |\beta|$) if $f(D(\beta|G)) \leqq f(D(\beta'|G))$ for all Y-elimination orderings β' of G; β is called *$\mathscr{F}_\ell$-minimal* if it is f-minimal for all $f \in \mathscr{F}_\ell$. To rank partial elimination orderings of a f-minimal elimination ordering the following functions

$$f_b(a_1, \cdots, a_m) := f(b_1, \cdots, b_r, a_1, \cdots, a_m) \quad \text{where}$$
$$b = (b_1, \cdots, b_r) \in \mathbb{N}^r, \quad r < n, \quad m = n - r$$

are derived from f. Evidently, f_b is a criterion function as defined in § 1, i.e., $f_b \in \mathscr{F}_m$. Furthermore we remark that

$$(3.2) \qquad f_b(a) = f_a(b) \quad \text{where } a = (a_1, \cdots, a_m) \in \mathbb{N}^m.$$

Using this notation we give an obvious and important identity: Let α be an elimination ordering of the graph $G = (X, E)$. Set $\alpha_1 := \alpha(1, r)$ and $\alpha_2 := \alpha(r + 1, n)$, where $r \in \mathbb{N}$, $1 \leqq r < n = |X|$. Then the following equalities hold:

$$(3.3) \quad f(D(\alpha|G)) = f(D(\alpha_1 + \alpha_2|G)) = f_{D(\alpha_1|G)}(D(\alpha_2|G_{\alpha_1})) = f_{D(\alpha_2|G_{\alpha_1})}(D(\alpha_1|G)).$$

The "kind of optimality" holding for partial elimination orderings of an f-minimal elimination ordering is now derived from (3.3).

PROPOSITION 3.4. *Given a graph $G = (X, E)$, $|X| =: n$, a criterion function $f \in \mathscr{F}_n$ and a f-minimal elimination ordering α of G. Set $\alpha_1 := \alpha(1, r)$, $\alpha_2 := \alpha(r + 1, n)$, $r \in \mathbb{N}$, $1 \leqq r < n$, and $Y := M(\alpha_1)$. Then it is true that*

(i) *α_1 is a $f_{D(\alpha_2|G_Y)}$-minimal Y-elimination ordering of G;*

(ii) *α_2 is a $f_{D(\alpha_1|G)}$-minimal elimination ordering of G_Y.*

Proof. (i) We assume that there is a Y-elimination ordering γ of G satisfying $f_{D(\alpha_2|G_Y)}(D(\gamma|G)) < f_{D(\alpha_2|G_Y)}(D(\alpha_1|G))$. According to (3.3), for the elimination ordering $\alpha' := \gamma + \alpha_2$ of G we get the inequality

$$f(D(\alpha'|G)) = f_{D(\alpha_2|G_Y)}(D(\gamma|G)) < f_{D(\alpha_2|G_Y)}(D(\alpha_1|G)) = f(D(\alpha|G))$$

which is a contradiction to the f-minimality of α. (ii) The proof is analogous to (i).

PROPOSITION 3.5. *In addition to the assumptions and notation of Proposition 3.4 let*

γ_1 be a $f_{D(\alpha_2|G_Y)}$-minimal Y-elimination ordering of G,

δ_1 be a $f_{D(\gamma_1|G)}$-minimal elimination ordering of G_Y,

δ_2 be a $f_{D(\alpha_1|G)}$-minimal elimination ordering of G_Y,

γ_2 be a $f_{D(\delta_2|G_Y)}$-minimal Y-elimination ordering of G.

Then the elimination orderings $\gamma_1 + \alpha_2$, $\gamma_1 + \delta_1$, $\alpha_1 + \delta_2$ and $\gamma_2 + \delta_2$ are f-minimal.

Proof. Using (3.3) we get

$$f(D(\alpha|G)) = f_{D(\alpha_2|G_Y)}(D(\alpha_1|G)) = f_{D(\alpha_2|G_Y)}(D(\gamma_1|G))$$
$$= f_{D(\gamma_1|G)}(D(\alpha_2|G_Y)) \geqq f_{D(\gamma_1|G)}(D(\delta_1|G_Y)) = f(D(\gamma_1 + \delta_1|G)).$$

Since α is f-minimal $\gamma_1 + \alpha_2$ and $\gamma_1 + \delta_1$ are f-minimal too. The f-minimality of $\alpha_1 + \delta_2$ and $\gamma_2 + \delta_2$ is proved analogously. □

Proposition 3.5 solves the notation problem of the introductory example: Given a graph $G = (X, E)$, $|X| =: n$, which is separable with respect to a criterion function $f \in \mathscr{F}_n$. Set (Y, S, Z) to be a decomposition of G with respect to f, $r := |Y|$ and α to be the corresponding f-minimal elimination ordering satisfying $\alpha(i) \in Y$ for $i = 1, \cdots, |Y|$, $\alpha(i) \in Z$ for $i = |Y| + 1, \cdots, |Y| + |Z|$ and $\alpha(i) \in S$ for $i = |Y| + |Z| + 1, \cdots, n$. According to Proposition 3.5 it is sufficient to determine first a $f_{D(\alpha(r+1,n)|G_Y)}$-minimal Y-elimination ordering γ_1 of G, and subsequently an $f_{D(\gamma_1|G)}$-minimal elimination ordering δ_1 of G_Y. Evidently, the elimination ordering $\gamma_1 + \delta_1$ is f-minimal. It is also possible to start with the determination of an $f_{D(\alpha(1,r)|G)}$-minimal elimination ordering δ_2 of G_Y. Subsequently, an $f_{D(\delta_2|G_Y)}$-minimal Y-elimination ordering γ_2 of G is required. In this situation $\gamma_2 + \delta_2$ is f-minimal. If G is small enough a combinatorical method for determining γ_1 and δ_1, respectively, γ_2 and δ_2, appears reasonable. But for an arbitrary $f \in \mathscr{F}$ a new problem arises. Since we do not really know the elimination ordering α the criterion functions $f_{D(\alpha(r+1,n)|G_Y)}$, respectively, $f_{D(\alpha(1,r)|G)}$, are unknown too. The criterion function given in Appendix A2 demonstrates this lack very clearly. Such kind of problems do not arise if we consider criterion functions like $f = f_L$ or $f = f_Q$ (which are interesting from the practical point of view) because in this situation minimizing $f_{D(\alpha(r+1,n)|G_Y)}$ is equivalent to minimizing $f_{(0,\cdots,0)}((0, \cdots, 0) \in \mathbb{N}^{n-r})$. This is easily verified by:

$$f_{D(\alpha(r+1,n)|G_Y)}(a_1, \cdots, a_r) = \sum_{i=1}^{r} a_i + \text{const},$$

where const $= \sum_{i=r+1}^{n} d(\alpha(i)|G_{\alpha(1,i-1)})$. Minimizing the function $f_{(0,\cdots,0)}$ is the formal description of the "partial minimization" used in the introductory example.

More interesting than computing γ_1 and δ_1, respectively, γ_2 and δ_2 by a combinatorical procedure is their determination using the Initial Theorem. This problem is considered for the rest of this section.

PROPOSITION 3.6. *Given a graph $G = (X, E)$, $|X| =: n$, along with a decomposition (Y, S, Z) of G with respect to $\mathscr{F}$. Furthermore set β to be a partial elimination ordering of G_Y satisfying the following two conditions*:

- *For any $g \in \mathscr{F}(G_Y)$: β may be chosen as a starting sequence of a g-minimal elimination ordering of G_Y*;
- *$M(\beta) \cap S = \varnothing$.*

Then for any $f \in \mathscr{F}$, β may be chosen as starting sequence of an f-minimal elimination ordering of G.

Proof. Let $f \in \mathscr{F}$, $|Y| =: r$. Set α to be a f-minimal elimination ordering satisfying $\alpha(i) \in Y$ for $i = 1, \cdots, r$, $\alpha(i) \in Z$ for $i = r + 1, \cdots, r + |Z|$ and $\alpha(i) \in S$ for $i = r + |Z| + 1, \cdots, n$ (which exists according to Definition 3.1). From the assumptions above it follows that β can be extended to an $f_{D(\alpha(1,r)|G)}$-minimal elimination ordering $\beta + \gamma$ of G_Y. Proposition 3.5 shows that $\alpha' := \alpha(1, r) + \beta + \gamma$ is f-minimal. Since $M(\gamma) \supset S$ Theorem 1.5 guarantees that $\alpha'' := \beta + \alpha(1, r) + \gamma$ is equivalent to α'. Therefore, α'' is f-minimal too. □

Since Definition 3.1 is symmetric with respect to Y and Z, it is possible that there are partial elimination orderings β_1 of G_Y and β_2 of G_Z both satisfying the assumptions of Proposition 3.6. If additionally S is complete in G_Y (or G_Z), then for any $f \in \mathscr{F}$, $\beta_1 + \beta_2$ may be chosen as a starting sequence of an f-minimal elimination ordering. The (short) proof of this statement is given in [We83].

In general G_Y contains more edges than $G(Z \dot{\cup} S)$, that edges which are introduced by the elimination of Y. Therefore, it may be possible that G_Y contains vertices of type B in contrast to $G(Z \dot{\cup} S)$. If at least one of these vertices is not contained in S, the partial elimination ordering β (required for Proposition 3.6) may be computed as a B-partial elimination ordering of G_Y. This is demonstrated by the following example.

Example. The graph $G = (X, E)$ (Fig. 4) does not contain any vertex of type B. Set $Y := \{y_1, y_2, y_3, y_4\}$, $\tilde{Z} := \{z_1, z_2, z_3, s_5\}$ and $S_1 := \{s_1, s_2, s_3, s_4\}$. Remark 2 on

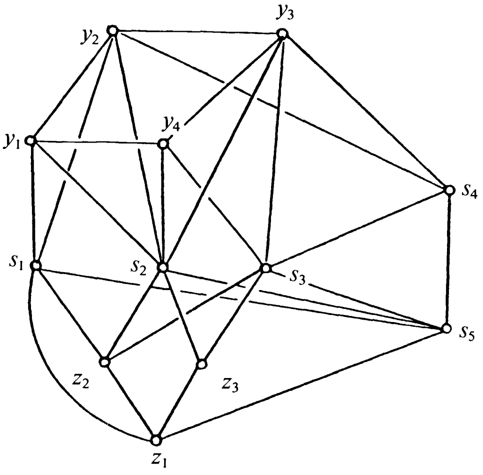


FIG. 4

Theorem 4.4 will guarantee that $(Y, S_1, \tilde{Z})$ is a decomposition of G with respect to $\mathscr{F}$. We consider the two elimination graphs G_Y and $G_{\tilde{Z}}$ (Fig. 5). For G_Y and $G_{\tilde{Z}}$ we determine two B-partial elimination orderings $\beta_1 := \langle z_3, z_2, z_1, s_5 \rangle$ (of G_Y) and $\beta_2 := \langle y_4, y_3, y_2, y_1 \rangle$ (of $G_{\tilde{Z}}$). Evidently, β_1 and β_2 satisfy the conditions of Proposition 3.6. Since S_1 is in G_Y (and $G_{\tilde{Z}}$) complete we have proved that

$$\alpha := \langle z_3, z_2, z_1, s_5, y_4, y_3, y_2, y_1, s_1, s_2, s_3, s_4 \rangle$$

is a $\mathscr{F}$-minimal elimination ordering of G.

A recursive application of this separation approach is also imaginable. It is feasible if (roughly speaking) the new edges which are introduced (in G_Y, respectively, G_Z) by the separation of G induces another separation of one (or both) of the graphs G_Y and G_Z. An example of this idea is given in [We83].

4. Separation theorems. In this section conditions on a graph G are established, sufficient for G to be separable with respect to a criterion function f. They depend on the graph G only (especially they are independent of the criterion function f) and therefore yield to decompositions of G with respect to $\mathscr{F}$.

The following definition generalizes the first assumption of the Initial Theorem ("the complementary graph of the neighbourhood of a type-B-vertex is a forest of bushes").

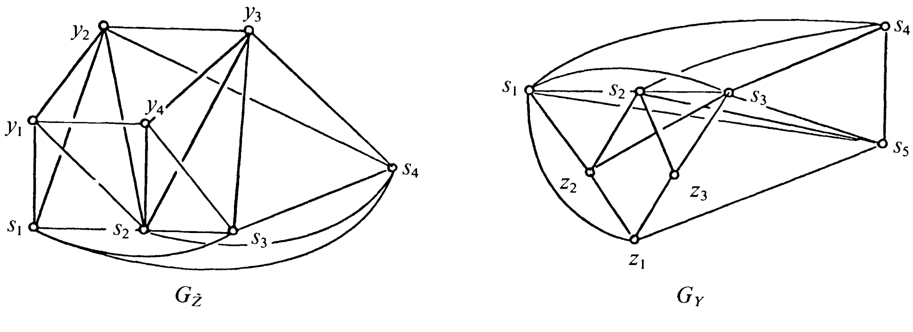


FIG. 5

DEFINITION 4.1. Let $F = (S, L)$ be a graph and $S_0 \subset S$ be complete. F is called *reducible to a complete graph by elimination of any k vertices* ($k \in \mathbb{N}$) *out of* $S \backslash S_0$ if

— $|S \backslash S_0| \geqq k$;

— The elimination graphs F_T are complete for all $T \subset S \backslash S_0$ with $|T| = k$.

This notation is used analogously for sets of vertices S of a graph $G = (X, E)$, i.e., S is reducible to a complete graph by elimination of any k vertices out of $S \backslash S_0$ ($S_0 \subset S$) if this is true for $G(S)$.

Remarks on Definition 4.1. (1) A graph $F = (S, L)$ whose complementary graph $\bar{F}$ is a forest of bushes is reducible to a complete graph by elimination of any vertex out of $S \backslash S_0$, where S_0 denotes the set of peaks of $\bar{F}$.

(2) Given a graph $F = (S, L)$ and a subgraph $\tilde{F} = (S, \tilde{L})$ of F ($\tilde{L} \subset L$). If F is reducible to a complete graph by elimination of any k vertices out of $S \backslash S_0$ ($S_0 \subset S$) then this holds true for F too. Attention has to be drawn to the following situation which is of particular interest in this paper: Given a graph $G = (X, E)$ and a set of vertices $S \subset X$ which is reducible to a complete graph by elimination of any k vertices out of $S \backslash S_0$ ($S_0 \subset S$ complete). Then this property remains true for any elimination graph G_T with $T \subset X \backslash S$.

The following two definitions state the assumptions for a first separation theorem.

DEFINITION 4.2. Given a graph $G = (X, E)$, $Y \subset X$, $S \subset X$ with $S \cap Y = \emptyset$ and $S_0 \subset S$ complete. The pair (S, S_0) is called of *type τ_k relative to Y* ($k \in \mathbb{N}$, $k \geqq 1$), if:

(i) $\operatorname{Adj}(Y|G) \subset S$.

(ii) For all $s, s' \in S$, $s \neq s'$, it is true:

$$(s, s') \notin E \Rightarrow \text{There are } k \text{ disjunct (nontrivial) paths } \omega_i, i = 1, \cdots, k, \text{ in } G \text{ from } s \text{ to } s' \text{ with } Z(\omega_i) \subset Y;$$

(iii) For any $y \in Y$ and to any $T \subset S \backslash S_0$ with $|T| = k$ there exist k disjunct paths ω_t, $t \in T$, from y to t with $Z(\omega_t) \subset Y$;

(iv) $|S \backslash S_0| \geqq k$.

We remark that (iii) and (iv) imply (evidently): For any $y \in Y$ and to any $T \subset S \backslash S_0$ with $|T| \leqq k$ there exist $|T|$ disjunct paths ω_t, $t \in T$, from y to t satisfying $Z(\omega_t) \subset Y$. Furthermore, condition (iv) is a technical one, which always holds true because Definition 4.2 is of interest only in connection with Definition 4.3.

DEFINITION 4.3. Given a graph $G = (X, E)$, $S \subset X$ and $S_0 \subset S$ complete. The pair (S, S_0) is called of *type T_k* ($k \in \mathbb{N}$, $k \geqq 1$) if the following conditions are satisfied:

(i) S is reducible to a complete graph by elimination of any k vertices out of $S \backslash S_0$.

(ii) S splits G into two graphs $G(Y)$ and $G(Z)$ (not necessarily connected), where $X \backslash S = Y \dot{\cup} Z$ (i.e., $G(X \backslash S) = G(Y) \oplus G(Z)$). The situation $Y = \emptyset$, respectively, $Z = \emptyset$ is, in contrast to the definition of a separating set of vertices (by technical reasons), also allowed.

(iii) (S, S_0) is of type τ_k relative to Y and relative to Z.

(iv) To any $s \in S \backslash S_0$ there are $m - 1$ ($m := |S|$) disjunct paths $\omega_{s'}$, $s' \in S$, $s' \neq s$, from s to s' with $Z(\omega_{s'}) \subset Z$.

Remarks on Definition 4.3. (1) If there is no doubt on the complete set of vertices S_0 we will also say that *S is of type T_k*.

(2) In order to employ a uniform notation we call (S, S_0) *of type* T_0, if conditions 4.3 (i) and (ii) are true; conditions 4.3 (iii)–(iv) are canceled. Evidently, sets of type T_0 are complete and separating. Their use to determine f-minimal elimination orderings is considered by Theorem 1.3. We remark that neither the elimination of Y nor the elimination of Z introduces new edges in $G(Z \cup S)$, respectively, $G(Y \cup S)$. Therefore, the method presented in § 3 (applying the Initial Theorem) will not succeed.

(3) Figure 6 illustrates two examples of sets of type T_k in its "basic form." Paths are indicated by ----.

(4) The conditions of Definitions 4.2 and 4.3 may be divided into two classes. The conditions 4.2 (i), (ii), (iv) and 4.3 (i), (iv) may be considered as "local with respect to S," because to verify these conditions we have to inspect only a (in general small) neighbourhood of S. In contrast the conditions 4.2 (iii) and 4.3 (ii) are global; their proof requires an inspection of the whole graph.

Using the two definitions introduced above we get a first separation theorem.

THEOREM 4.4. *Given a graph $G = (X, E)$ containing a set S of vertices of type T_k, then for any $f \in \mathcal{F}$ there exists an f-minimal elimination ordering α of G (depending on f) which satisfies*

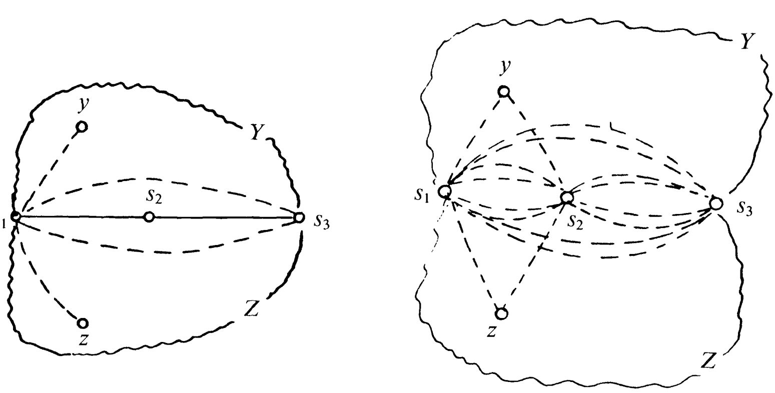
$$\alpha(i) \in \begin{cases} Y & \text{for } i = 1, \cdots, |Y|, \\ Z & \text{for } i = |Y| + 1, \cdots, |Y| + |Z|, \\ S & \text{for } i = |Y| + |Z| + 1, \cdots, |X|. \end{cases}$$

The notation Y, Z and S is used in the sense of Definition 4.3.

A *proof* of Theorem 4.4 is given in Appendix A4.

Remarks on Theorem 4.4. (1) The feasibility of Theorem 4.4 depends essentially on the chance of finding sets of vertices of type T_k. This problem may be considered from two distinct points of view. On the one hand, if we are interested in an algorithmic search, no efficient procedure for solving this problem can be expected. This depends essentially on two facts: First, the verification of Definitions 4.2 and 4.3 is a very expensive task. Second, a lot of subsets of vertices (all of X) must be checked. On the other hand in a "clear" graphical representation of a graph sets of vertices of type T_k can be recognized frequently (if contained) because Definitions 4.2 and 4.3 are based on the visual notions "separating" and "disjunct paths." In this context it should be noticed that in some fields of applications clear graphical representations of the considered graphs are available (for example the network graph in the field of load-flow calculation in power systems). We summarize: Theorem 4.4 does not lead to an efficient programmable algorithm for the determination of f, respectively, $\mathcal{F}$ minimal elimination orderings but it is a useful tool for attacking this problem by working with "paper and pencil."

(2) In [We83] a more general separation theorem is proved. The definition of the property T_k must be generalized as follows.



$k = 1$,
$S = \{s_1, s_2, s_3\}$,
$S_0 = \{s_2, s_3\}$.

$k = 2$,
$S = \{s_1, s_2, s_3\}$,
$S_0 = \{s_3\}$.

FIG. 6

DEFINITION 4.3′. Given a graph $G = (X, E)$, $S \subset X$ and $S_0 \subset S$ complete. The pair (S, S_0) is called *of type* T_k ($k \in \mathbb{N}$, $k \geqq 1$), if the following conditions are satisfied:

(i) Analogously Definition 4.3 (i) (remains unchanged);

(ii) Analogously Definition 4.3 (ii) (remains unchanged);

(iii) There are subsets $S_1 \subset S$ and $S_2 \subset S$ satisfying

— $S = S_1 \cup S_2$,

— $S \backslash S_0 \subset S_1 \cap S_2$,

— $m_1 := |S_1| \leqq |S_2| =: m_2$.

For S_1 and S_2 the following conditions (iv) and (v) also hold.

(iv) $(S_1, S_1 \cap S_0)$ is of type τ_k relative to Y. $(S_2, S_2 \cap S_0)$ is of type τ_k relative to Z.

(v) For any $s \in S \backslash S_0$ there are $m_2 - 1$ disjunct paths $\omega_{s'}$, $s' \in S_2$, $s' \neq s$, from s to s' with $Z(\omega_{s'}) \subset Z$.

Using this notation, [We83, 8.2.4] guarantees that for a graph G containing a set S of vertices of type T_k (in the sense defined above) there exists an f-minimal elimination ordering α (depending on $f \in \mathcal{F}$) which satisfies

$$\alpha(i) \in \begin{cases} Y & \text{for } i = 1, \cdots, |Y|, \\ \tilde{Z} & \text{for } i = |Y| + 1, \cdots, |Y| + |\tilde{Z}|, \\ S_1 & \text{for } i = |Y| + |\tilde{Z}| + 1, \cdots, |X|, \end{cases}$$

where $\tilde{Z} := Z \cup (S_2 \backslash S_1)$.

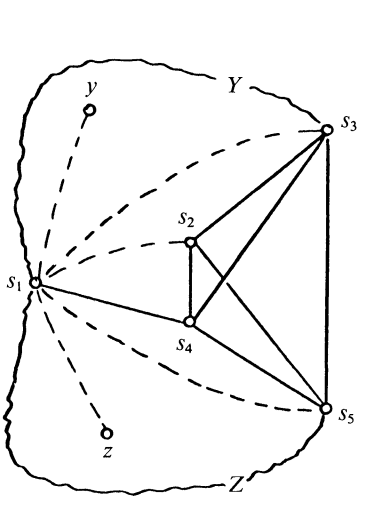
Figure 7 gives an example of a set of type T_k (in the sense of Definition 4.3′); paths are indicated by ---- again.

(3) Counterexamples demonstrate that 4.2 (ii) and 4.3 (iv) may not be canceled, by the assumption of 4.4 [We83, 8.2.4, Remark 4].

The following example demonstrates the application of Theorem 4.4 in connection with the Initial Theorem.

Example. We consider the graph $G = (X, E)$, shown in Fig. 8. G is interesting from the practical point of view. It is the elimination graph (after repeated elimination of vertices of type B) of a graph containing 175 vertices originally representing an electrical power system (a connection of the AEP-118-node test network by the AEP-57-node test network; see [We83]). In G

$$S := \{123, 128, 129\}$$



$k = 1$,
$S_1 = \{s_1, s_2, s_3\}$,
$S_2 = \{s_1, s_4, s_5\}$,
$S_0 = \{s_2, s_3, s_4, s_5\}$.

FIG. 7

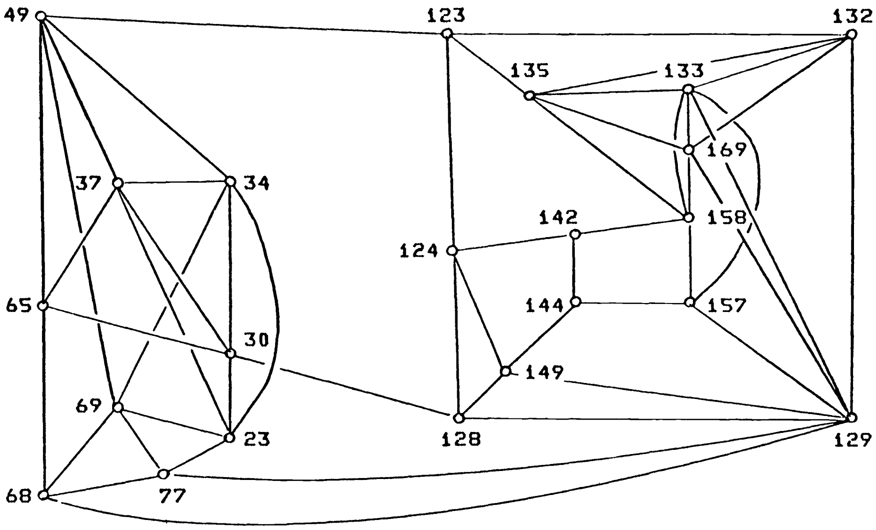


FIG. 8

is of type T_1, where $S_0 := \{128,\ 129\}$. For Y and Z we get:

$$Y = \{23, 30, 34, 37, 49, 65, 68, 69, 77\} \quad \text{and}$$

$$Z = \{124, 135, 132, 133, 169, 158, 157, 142, 144, 149\}.$$

Now we consider (as described in § 3) the graph $G' := G_Y$ (Fig. 9). It is easily verified that

$$\beta' := \langle 132, 135, 169, 133 \rangle$$

is a B-partial elimination ordering of G' with $M(\beta') \cap S = \emptyset$. Therefore, for any $f \in \mathscr{F}(G)$, β' is a starting sequence of an f-minimal elimination ordering of G. The graph $G'' := G_{\beta'}$ (Fig. 10) remains to be investigated. G'' does not contain any vertex of type B.

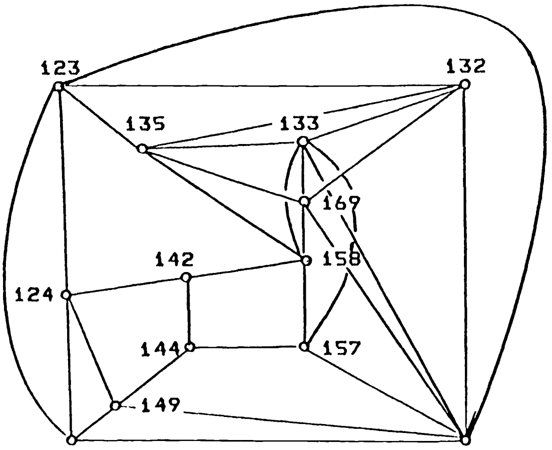


FIG. 9

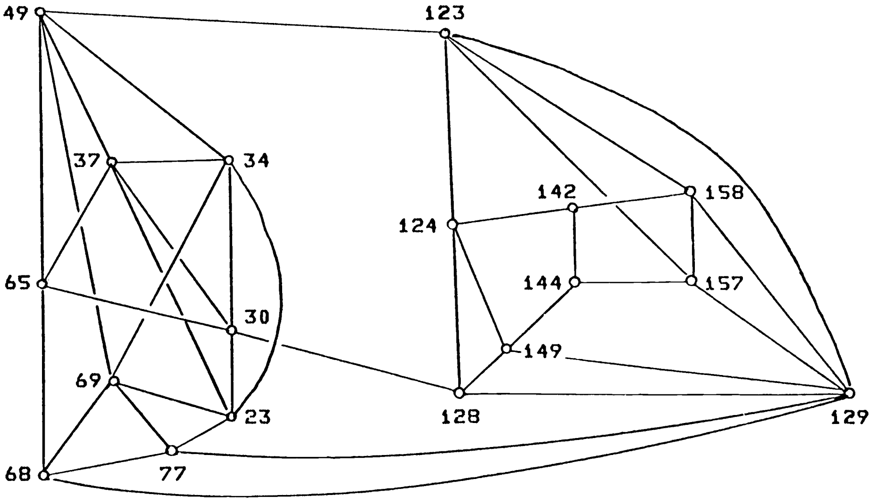


FIG. 10

We apply Theorem 4.4 again. In G'' the set of vertices

$$S'' := \{123, 124, 149, 129\}$$

is of type T_1, where $S_0 := \{149, 129\}$. For Y'' resp. Z'' we get

$$Y'' = Y \cup \{128\}, \qquad Z'' = \{144, 142, 158, 157\}.$$

Now we have to look at the graph $G^{(3)} := G''_{Y''}$ (Fig. 11). Taking advantage of the symmetry of $G^{(3)}$ we see: To any $f \in \mathscr{F}(G^{(3)})$ there exists an f-minimal elimination ordering (depending on f) starting with vertex 142 [We83, Example 7.3]. For $G^{(3)}_{\{142\}}$ we compute the B-partial elimination ordering $\gamma := \langle 157, 158, 144 \rangle$. Hence, we have proved: For any $f \in \mathscr{F}(G^{(3)})$

$$\beta'' := \langle 142, 157, 158, 144 \rangle$$

is a starting sequence of an f-minimal elimination ordering of $G^{(3)}$ which satisfies $M(\beta'') \cap S'' = \emptyset$. Thus, for any $f \in \mathscr{F}(G'')$, β'' is a starting sequence of an f-minimal elimination ordering of G''. The remaining graph $G^{(4)} := G''_{\beta''}$ (Fig. 12) allows a B-elimination ordering:

$$\alpha'' := \langle 124, 149, 123, 128, 65, 129, 34, 37, 23, 30, 49, 68, 69, 77 \rangle.$$

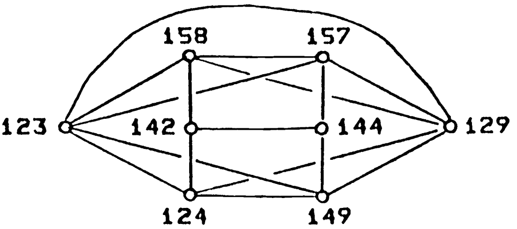


FIG. 11

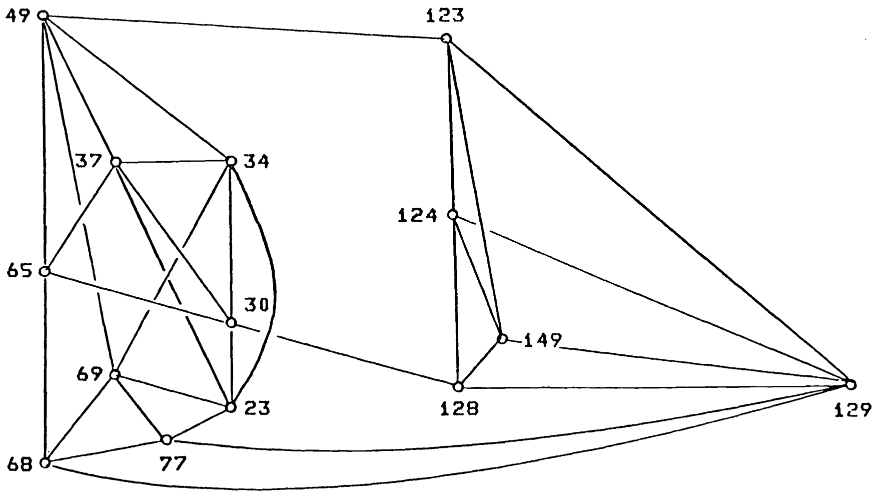


FIG. 12

Summarizing, we have proved that

$$\begin{aligned}\alpha &:= \beta' + \beta'' + \alpha'' \\ &= \langle 132, 135, 169, 133, 142, 157, 158, 144, 124, 149, 123, 128, 65, 129, 34, \\ &\quad 37, 23, 30, 49, 68, 69, 77\rangle\end{aligned}$$

is an $\mathscr{F}$-minimal elimination ordering of G.

A further interesting example is the grid graph $G = (X, E)$ of Fig. 13. An $\mathscr{F}$-minimal elimination ordering of G is determined in [We83].

If $k = 1$ condition 4.2 (iii) may be canceled from the assumptions of Theorem 4.4. A set of vertices S is called *almost of type* T_1 if it is type T_1 except for condition 4.2 (iii).

THEOREM 4.5. *Given a graph* $G = (X, E)$ *containing a set of vertices* S *which is almost of type* T_1. *Then to any* $f \in \mathscr{F}$ *there exists an f-minimal elimination ordering* α *of* G (*depending on* f) *which satisfies*

$$\alpha(i) \in \begin{cases} Y & \textit{for } i = 1, \cdots, |Y|, \\ Z & \textit{for } i = |Y|+1, \cdots, |Y|+|Z|, \\ S & \textit{for } i = |Y|+|Z|+1, \cdots, |X|. \end{cases}$$

The notation Y and Z are chosen according to Definition 4.3, *respectively, Theorem* 4.4.

A proof of Theorem 4.5 is given in [We83].

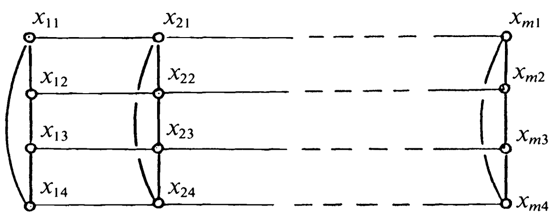


FIG. 13

It is interesting to consider Theorem 4.5 for the special case $|Y| = 1$. It is easily verified that the neighbourhood of a vertex of type B is almost of type T_1. Therefore, the Initial Theorem is a straightforward corollary of Theorem 4.5. The Initial Theorem is not involved in the proof of Theorem 4.5.

Now we turn to a third separation theorem which is weakening condition 4.2 (ii) for $k = 2$. Its assumptions are stated in two definitions again.

DEFINITION 4.6. Given a graph $G = (X, E)$, $Y \subset X$, $S \subset X$ with $S \cap Y = \emptyset$, $S_0 \subset S$ complete and $|S \backslash S_0| = 2$. Set $\{s_1, s_2\} := S \backslash S_0$, $s_1 \neq s_2$. The pair (S, S_0) is called *of type τ'_2 relative to Y* if:

(i) $\mathrm{Adj}\,(Y|G) \subset S$.

(ii) For all $s, s' \in S$, $s \neq s'$, it is true:

$$(s, s') \notin E \Rightarrow \text{There exist two disjunct (nontrivial) paths } \omega_i, i = 1, 2, \text{ from } s \text{ to } s' \text{ satisfying } Z(\omega_i) \subset Y \cup \{s_1, s_2\}.$$

(iii) For all $y \in Y$ there exist two disjunct paths ω_i, $i = 1, 2$, from y to s_i satisfying $Z(\omega_i) \subset Y$.

DEFINITION 4.7. Given a graph $G = (X, E)$, $S \subset X$, $S_0 \subset S$ complete and $|S \backslash S_0| = 2$. Set $\{s_1, s_2\} := S \backslash S_0$, $s_1 \neq s_2$. The pair (S, S_0) is called *of type T'_2* if the following conditions are satisfied:

(i) S splits G into two graphs $G(Y)$ and $G(Z)$ (not necessarily connected), where $X \backslash S = Y \dot{\cup} Z$ (i.e., $G(X \backslash S) = G(Y) \oplus G(Z)$). The situation $Y = \emptyset$, respectively, $Z = \emptyset$ is, in contrast to the definition of a separating set of vertices (by technical reasons) also allowed.

(ii) (S, S_0) is of type τ'_2 relative to Y and relative to Z.

(iii) For s_i, $i = 1, 2$, there exist $m - 1$ ($m := |S|$) disjunct paths $\omega_{s'}$, $s' \in S$, $s' \neq s_i$, from s_i to s' satisfying $Z(\omega_{s'}) \subset Z$.

If there is no doubt on the complete set S_0 we simply speak of S to be of type T'_2. Figure 14 illustrates sets of type T_2; paths are indicated by ---.

We get the third separation theorem as follows.

THEOREM 4.8. *Given a graph $G = (X, E)$ containing a set of vertices $S \subset X$ of type T'_2, for each $f \in \mathcal{F}$ there exists an f-minimal elimination ordering α of G (depending on f) which satisfies*

$$\alpha(i) \in \begin{cases} Y & \textit{for } i = 1, \cdots, |Y|, \\ Z & \textit{for } i = |Y| + 1, \cdots, |Y| + |Z|, \\ S & \textit{for } i = |Y| + |Z| + 1, \cdots, |X|. \end{cases}$$

A proof of Theorem 4.8 is given in [We83].

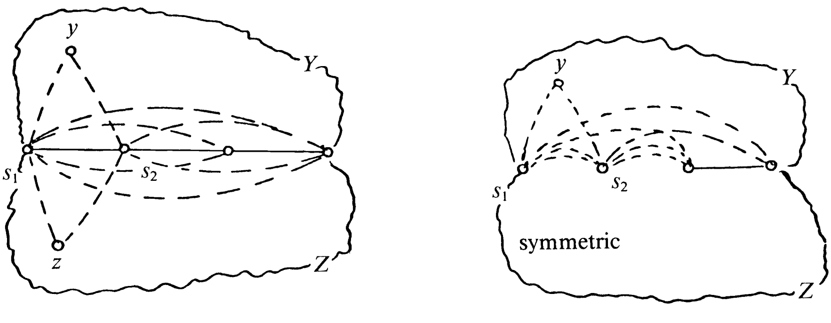
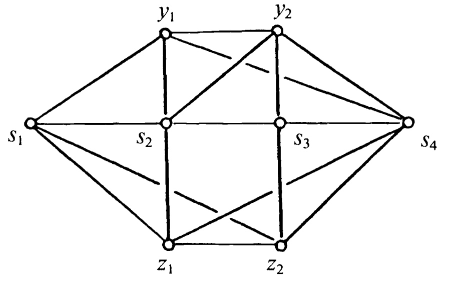


FIG. 14



$S = \{s_1, s_2, s_3, s_4\}$,
$S_0 = \{s_3, s_4\}$.

FIG. 15

Examples. (i) Consider the graph of Fig. 15. Obviously, (S, S_0) is the type T'_2. As described in § 3, we get an $\mathscr{F}$-minimal elimination ordering

$$\alpha = \langle y_1, y_2, z_1, z_2, s_1, s_2, s_3, s_4 \rangle$$

of G.

(ii) In [We83], for the grid graph $G = (X, E)$ (Fig. 16), it is proved that the elimination ordering

$$\alpha = \beta' + \beta'' + \alpha''$$

is $\mathscr{F}$-minimal, where

$$\beta' := \langle x_{11}, x_{14}, x_{12}, x_{21}, x_{24}, x_{m1}, x_{m4}, x_{m2}, x_{(m-1)1}, x_{(m-1)4} \rangle;$$

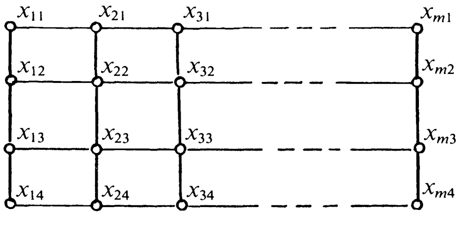
$$\beta'' := \langle x_{13}, x_{22}, x_{23} \rangle;$$

$$\alpha'' := \langle x_{32}, x_{33}, x_{41}, x_{44}, x_{31}, x_{34}, x_{42}, x_{43}, \cdots, x_{(m-3)3}, x_{m3}, x_{(m-1)2},$$

$$x_{(m-1)3}, x_{(m-2)1}, \cdots, x_{(m-2)4} \rangle, \text{ where } m \text{ is assumed to be odd.}$$

If m is even a similar B-elimination ordering can be computed.

5. f-minimal elimination orderings of simplex graphs. In this section we consider the problem of determining f, respectively, $\mathscr{F}$-minimal, elimination orderings for a special type of graphs, called simplex graphs. Simplex graphs are of interest in the field of load flow calculation in power systems. In general the (nonlinear) network equations are solved by means of Newton's method, which involves a repetitive solving of a system of linear equations, the Jacobi matrix equation [St74]. During the (repetitive) solution process the zero-nonzero pattern of the matrix remains fixed; only its entries change. In general those systems are not positive definite but large experience shows that diagonal pivots are appropriate from the numerical point of view. Since the Jacobi matrix is symmetric in its zero-nonzero pattern, the graph-theoretical model of the Gaussian elimination



$m \geqq 6$

FIG. 16

process is feasible. The zero-nonzero pattern of the Jacobi matrix is represented by a simplex graph. Simplex graphs arise also in the field of structural analysis [Sch80].

DEFINITION 5.1. A graph $G = (X, E)$ is called a *simplex graph*, if there is a partition $X = X_1 \dot{\cup} \cdots \dot{\cup} X_n$, $n \in \mathbb{N}$, of the set of vertices which satisfies:

(i) $G(X_i)$ is complete for $i = 1, \cdots, n$.

(ii) If $\operatorname{Adj}(X_i|G) \cap X_j \neq \emptyset$ for $i \neq j$ then $G(X_i \cup X_j)$ is complete.

The sets X_i are called *blocks* of G, the number $\dim(X_i) := |X_i|$ is denoted as *dimension* of the block X_i. If $\dim(X_i) = r$ for all $i = 1, \cdots, n$, we call G *r-simplex graph.*

Evidently, every graph is a 1-simplex graph. Furthermore it is easily verified that an r-simplex graph is a r'-simplex graph if r' divides r. But we are not interested in that kind of question and assume for the rest of this section that the partition belonging to a simplex graph is always fixed and well known; especially, all notation and considerations are referred to that partition.

Another method to introduce simplex graphs is a constructive one. It involves the 'global structure' inherent in a simplex graph. In this section, for a graph $G = (X, E)$ along with a valuation d of its vertices (meaning a map $d : X \mapsto \mathbb{N}$), we write $G = (X, E, d)$.

DEFINITION 5.2. Let $G = (X, E, d)$ be a graph with a valuation of its vertices. A simplex graph $G^* = (X^*, E^*)$ is derived from G in the following manner:

— For each $x \in X$ let x^* be a set of $d(x)$ distinct vertices; for two distinct vertices $x, y \in X$ the corresponding sets x^* and y^* are defined to be disjunct. The total of the vertices of G^* becomes $X^* := \cup_{x \in X} x^*$.

— The set of edges E^* of G^* is defined by: For any $u, v \in X^*$, $u \neq v$, with $u \in x^*$ and $v \in y^*$ we set:

$$(u, v) \in E^* :\Leftrightarrow (x = y \vee (x, y) \in E).$$

G^* is called *simplex extension of* G. Evidently, $\dim(x^*) = d(x)$ for all $x \in X$. If $d(x) = r$ for all $x \in X$, we call G^* the *r-simplex extension of* G; in this case the valuation d is omitted.

Intuitively, the simplex extension of a graph $G = (X, E, d)$ is constructed by

— Replacing the vertices $x \in X$ by cliques of size $d(x)$;

— "Connecting two distinct cliques completely" if and only if the corresponding vertices are adjacent.

See Fig. 17 for an example. Obviously, every simplex graph $\tilde{G}$ is (respectively, can be considered as) a simplex extension G^* of a graph $G = (X, E, d)$ with a suitable valuation

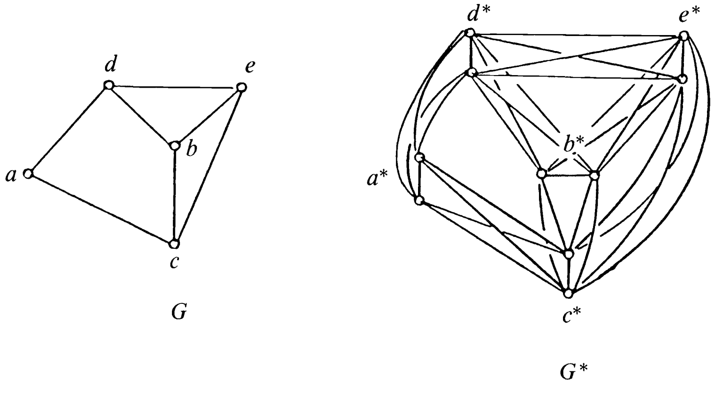


FIG. 17

d of its vertices. In this context G is called the *skeleton of* $\tilde{G}$. Therefore, it is sufficient to consider simplex extensions only.

We are interested in f, respectively, $\mathscr{F}$-minimal elimination orderings of simplex graphs. Attacking this problem by means of § 2, we see that the Initial Theorem is not feasible because generally simplex graphs do not contain vertices of type B. This is easily demonstrated by the small example given above (Fig. 17). In the following a more general Initial Theorem, appropriate for simplex graph, will be developed. But first we provide some notations and "rules of computation" for handling simplex extensions.

PROPOSITION 5.3. *For a graph* $G = (X, E, d)$ *and the corresponding simplex extension* $G^* = (X^*, E^*)$,

$$(G_x)^* = (G^*)_{x^*} \quad \textit{for all } x \in X.$$

Proof. The proof is obvious.

DEFINITION 5.4. Given a graph $G = (X, E, d)$, $|X| =: n$, along with the corresponding simplex extension $G^* = (X^*, E^*)$, $n^* := |X^*|$. An elimination ordering α of G^* is called *eliminating block by block* if the following condition is satisfied for $1 \leqq j \leqq n^*$:

$$\alpha(j) \in x^* \wedge M(\alpha(j+1, n^*)) \cap x^* \neq \varnothing \Rightarrow \alpha(j+1) \in x^*.$$

Every elimination ordering α eliminating block by block can be represented by $\alpha = \beta_1 + \cdots + \beta_n$, where β_i denotes a partial elimination ordering which eliminates exactly one block of G^*. Furthermore we remark that for any $f \in \mathscr{F}(G^*)$ the number $f(D(\alpha|G^*))$ does not depend on the internal order of elimination of the β_i. We prove this property for $i = 1$ only; the general case is easily concluded by induction. Let $\alpha(1) \in x^* := \{v_1, \cdots, v_c\}$, $c := \dim(x^*)$ and without loss of generality set $\alpha(1, c) = \beta_1 = \langle v_1, \cdots, v_c \rangle$. By definition, for a simplex graph, $d(v_1|G^*) = d(v_i|G^*)$ for $i = 1, \cdots, c$. Furthermore we get $d(v_2|G^*_{\langle v_1 \rangle}) = d(v_1|G^*) - 1$, $d(v_3|G^*_{\langle v_1, v_2 \rangle}) = d(v_1|G^*) - 2$, and so on until $d(v_c|G^*_{\langle v_1, \cdots, v_{c-1} \rangle}) = d(v_1|G^*) - (c - 1)$. This shows that α is equivalent to each elimination ordering $\langle v_{\pi(1)}, \cdots, v_{\pi(c)} \rangle + \alpha(c+1, n^*)$, where π denotes any permutation of the numbers $1, \cdots, c$. According to the property proved above any elimination ordering of G^* which eliminates block by block can be interpreted as an "extension" α^* of an elimination ordering α of G. More exactly, α^* is defined by $\alpha^* := \langle \alpha(1)^* \rangle + \cdots + \langle \alpha(n)^* \rangle$ of G^*, where $\langle \alpha(i)^* \rangle$ denotes any fixed ordering of the vertices $\alpha(i)^*$, yet this ordering is not of interest in detail.

In order to determine an f-minimal elimination ordering of a simplex graph it is sufficient to consider only elimination orderings which eliminate block by block. This is guaranteed by the following proposition.

PROPOSITION 5.5. *Let* $G = (X, E, d)$ *be a graph with a valuation of the vertices,* $G^* = (X^*, E^*)$ *is the corresponding simplex extension. Then, to any elimination ordering* α *of* G^* *there exists an elimination ordering* α' *of* G^* *which dominates* α *and which eliminates block by block.*

Proof. Let x^* be that block of G^* which contains $\alpha(1)$. According to the definition of a simplex graph, Adj $(x^* \backslash \{\alpha(1)\}|G^*) \cup (x^* \backslash \{\alpha(1)\})$ is complete in $G^*_{\alpha(1)}$. Therefore, there exists an elimination ordering of α' of G^* which dominates α and which satisfies $M(\alpha'(1, d(x))) = x^*$. The proof is accomplished by induction. □

Now we state an initial theorem for simplex graphs.

DEFINITION 5.6. Given a simple extension $G^* = (X^*, E^*)$ along with the corresponding graph $G = (X, E, d)$. A block x^* of G^* is called *of type* B* if the following conditions hold:

(i) $\overline{G(\text{Adj}\,(x|G))}$ is a forest of bushes. We set W to be the set of roots, B to be the set of peaks and $B(w)$, $w \in W$, to be the set of peaks belonging to the root w.

(ii) $d(w) = \dim(w^*) \leqq \dim(x^*) = d(x)$ for all $w \in W$.

(iii) For each root $w \in W$ it is true: For at least one $v \in w^*$ there are disjunct paths ω_c $c \in \bigcup_{b \in B(\omega)} b^*$, in G^* from v to c with $Z(\omega_c) \cap (\text{Adj}\,(x^*|G^*) \cup x^*) = \varnothing$.

THEOREM 5.7. *Given a simplex extension $G^* = (X^*, E^*)$ along with the corresponding graph $G = (X, E, d)$ and a block x^* of G^* of type* B*. *Then to each elimination ordering α' of G^* there exists an elimination ordering α of G^* which dominates α' and which satisfies $M(\alpha(1, d(x))) = x^*$ (meaning that the block x^* is eliminated at the beginning of α).*

Theorem 5.7 yields the next theorem directly.

THEOREM 5.8 (Initial theorem for simplex graphs). *Given a simplex extension $G^* = (X^*, E^*)$ along with the corresponding graph $G = (X, E, d)$ and a block x^* of G^* of type* B*. *Then to each $f \in \mathcal{F}(G^*)$ there exists an f-minimal elimination ordering α of G^* (depending on f) with $M(\alpha(1, d(x)) = x^*$.*

The proof of 5.7 is given in Appendix A5.

Remarks on Theorem 5.8. (1) If condition 5.6 (iii) is satisfied for one $v \in w^*$ it is satisfied for all $v \in w^*$.

(2) In [We83] f, respectively, $\mathcal{F}$, minimal elimination orderings are determined for a lot of simplex graphs using Theorem 5.7 only. The examples have been chosen from the field of load-flow calculation in power systems. The problem of finding disjunct paths (condition 5.6 (iii)) has been treated as a flow-problem.

(3) The following counterexample demonstrates that the additional condition 5.6 (ii) (compared to Definition 2.1) cannot be removed from the assumptions of Theorem 5.8. Consider the simplex extension G^* of G (Fig. 18). We compare the two elimination orderings $\alpha^* = \langle a^*, b^*, c^*, d^* \rangle$ and $\beta^* = \langle b^*, a^*, c^*, d^* \rangle$ of G^*. The elimination ordering α^* eliminates successively blocks of type B*. In contrast, for β^* condition 5.6 (ii) does not hold in the first step of elimination. Calculating $f_L(D(\alpha^*|G^*)) = 4 + 3 + 4 + 3 + 2 + 1 = 17$ and $f_L(D(\beta^*|G^*)) = 4 + 5 + 4 + 3 + 2 + 1 = 19$ shows that β^* is not f_L-minimal.

(4) For an r-simplex extension G^* of G the following two statements are equivalent:

— x is of type B in G;

— x^* is of type B* in G^*.

(5) By a suitable interpretation of 5.6 (iii) the property of a block to be of type B* can be defined in terms of the generating graph $G = (X, E, d)$. Condition 5.6 (iii) must be modified to:

For each root $w \in W$, for each $b \in B(w)$ there exist paths $\omega_b^{(i)}$, $i = 1, \cdots, d(b)$ (not necessarily disjunct) from w to b with $Z(\omega_b^{(i)}) \cap (\text{Adj}\,(x|G) \cup \{x\}) =$

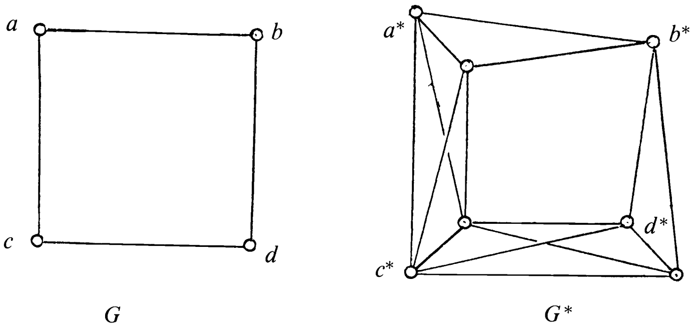


FIG. 18

$\varnothing$. Furthermore the totality of these paths $\{\omega_b^{(i)} | b \in B(w),\ 1 \leqq i \leqq d(b)\}$ satisfies: Every vertex $z \in X \backslash (\mathrm{Adj}\,(x|G) \cup \{x\})$ belongs to at most $d(z)$ of these paths.

This condition guarantees that the corresponding simplex extension contains the paths claimed by condition 5.6 (iii).

The rest of this section deals with the determination of f, respectively, $\mathscr{F}$, minimal elimination orderings of r-simplex extensions. This problem can be treated by considering the corresponding skeleton only.

PROPOSITION 5.9. *Given a graph $G = (X, E)$, $n := |X|$, along with the corresponding r-simplex extension $G^* = (X^*, E^*)$. Furthermore let α be any elimination ordering of G; α^* is the corresponding r-simplex extension of α. Then for all $f \in \mathscr{F}(G^*)$,*

$$f(D(\alpha^*|G^*)) = (f \times \nu)(D(\alpha|G)),$$

where

$$\nu : \mathbb{N} \to \mathbb{N}^r, \qquad \nu(a) = (r \cdot a + r - j)_{j=1,\cdots,r}$$

and

$$f \times \nu : \mathbb{N}^n \to \mathbb{R}, \qquad (f \times \nu)(a_1, \cdots, a_n) := f(\nu(a_1), \cdots, \nu(a_n)).$$

Especially: $f \times \nu \in \mathscr{F}(G)$.

Proof. It is easily verified that the equality $d(x^*|G^*) = (r \cdot d(x|G) + r - j)_{j=1,\cdots,r}$ holds true for any block x^* of G^*. Thus

$$\begin{aligned} f(D(\alpha^*|G^*)) &= f(d(\alpha(1)^*|G^*), \cdots, d(\alpha(n)^*|G^*_{\alpha(1,n-1)^*})) \\ &= f(((r \cdot d(\alpha(i)|G_{\alpha(1,i-1)}) + r - j)_{j=1,\cdots,r})_{i=1,\cdots,n}) \\ &= (f \times \nu)(D(\alpha|G)). \end{aligned}$$

Therefore, the first part of Proposition 5.9 is proved. Since $r > 0$ we have: $a \leqq b \Rightarrow \nu(a) \leqq \nu(b)$ (for each entry), respectively, $(a_1, \cdots, a_n) \leqq (b_1, \cdots, b_n) \Rightarrow f(\nu(a_1), \cdots, \nu(a_n)) \leqq f(\nu(b_1), \cdots, \nu(b_n))$. The symmetry of criterion functions implies the symmetry of $f \times \nu$. Altogether it is proved that $f \times \nu$ is a criterion function. □

Remark on Proposition 5.9. For readers who are interested in the special criterion function f_L and f_Q, we calculate $f_L \times \nu$ and $f_Q \times \nu$ explicit:

$$f_L \times \nu = r^2 \cdot \tilde{f}_L + n \cdot r \cdot (r-1)/2, \qquad (f_L f_Q \in \mathscr{F}(G^*), \tilde{f}_L \tilde{f}_Q \in \mathscr{F}(G)),$$

$$f_Q \times \nu = r^3 \cdot \tilde{f}_Q + r^2 \cdot (r-1) \cdot \tilde{f}_L + n \cdot (r-1) \cdot r \cdot (2r-1)/6.$$

From Proposition 5.9 we get the following.

COROLLARY 5.10. *If we use the notation of Proposition* 5.9 *then it is true that*

(i) α *($f \times \nu$)-minimal* $\Leftrightarrow$ α^* *f-minimal*;

(ii) α *$\mathscr{F}(G)$-minimal* $\Rightarrow$ α^* *$\mathscr{F}(G^*)$-minimal.*

According to Corollary 5.10 the determination of an f, respectively, $\mathscr{F}$ minimal, elimination ordering of a simplex graph $\tilde{G}$ is reduced to the construction of an ($f \times \nu$)-minimal, respectively, $\mathscr{F}(G)$-minimal, elimination ordering of a skeleton of $\tilde{G}$. There are several methods to attack this problem. In this context a question arises which is interesting especially from the practical point of view: Are there two distinct skeletons of an r-simplex graph, one which is suited for the determination of a f-minimal elimination ordering and the other which is not? From the example given at the beginning of this section, this situation appears quite possible. But the following proposition guarantees that the skeleton is an invariant of a r-simplex graph.

PROPOSITION 5.11. *All skeletons of an r-simplex graph are isomorphic to each other.*
Proof. A proof is given in [We83].

Appendix A1. Definitions, statements and "roles of computation" for handling elimination graphs. Let $G = (X, E)$ be a graph with $|X| = n$. By definition, for each sequence $\langle x_1, \cdots, x_k\rangle$ of distinct vertices the elimination graph $G_{\langle x_1, \cdots, x_k\rangle}$ is well defined. In [Be72, Proof of Thm. 2.6.1] it is shown that, for the edges (a, b) belonging to $G_{\langle x_1, \cdots, x_k\rangle}$ $(a, b \in X \backslash \{x_1, \cdots, x_k\}, a \neq b)$,

$$\text{(A1.1)} \qquad (a, b) \in E_{\langle x_1, \cdots, x_k\rangle} \Leftrightarrow \text{There exists a path } \omega \text{ in } G \text{ from } a \text{ to } b \text{ with } Z(\omega) \subset \{x_1, \cdots, x_k\}.$$

Therefore, the elimination graph $G_{\langle x_1, \cdots, x_k\rangle}$ does not depend on the order of elimination of the vertices $x_1, \cdots, x_k$. Consequently, for any set $A = \{a_1, \cdots, a_k\}$ of distinct vertices of G the elimination graph $G_A = (X_A, E_A) := G_{\langle a_1, \cdots, a_k\rangle}$ is well defined. G_A is called the A-*elimination graph* of G. From (A1.1) it follows that for $x, y \in X \backslash A$, $x \neq y$,

$$\text{(A1.2)} \qquad (x, y) \in E_A \Rightarrow (x, y) \in E \vee (x \in \text{Adj}\,(A|G) \wedge y \in \text{Adj}\,(A|G)).$$

Another trivial but useful conclusion from (A1.1) is

(A1.3) For every connected set $A \subset X$ (i.e., $G(A)$-connected) the neighbourhood Adj $(A|G)$ is complete in G_A (i.e., $G_A(\text{Adj}\,(A|G))$-complete).

The modification of the neighbourhood of a vertex $x \in X$ caused by the elimination of a set of vertices $A \subset X$, $x \notin A$, is derived from (A1.1) too [Be72, Thm. 2.6.3]:

$$\text{(A1.4)} \qquad \text{Adj}\,(x|G_A) = \text{Adj}\,(x|G) \backslash A \cup \bigcup_{\substack{i=1 \\ x \in \text{Adj}\,(A_j|G)}}^{n} \text{Adj}\,(A_j|G) \backslash \{x\}.$$

where the sets of vertices of the (connected) components of $G(A)$ are denoted by A_j. For $A = \{y\}$ (and $y \neq x$) (A1.4) reduces to

$$\text{(A1.5)} \qquad \text{Adj}\,(x|G_y) = \begin{cases} \text{Adj}\,(x|G) & \text{for } (x, y) \notin E, \\ \text{Adj}\,(x|G) \backslash \{y\} \cup \text{Adj}\,(y|G) \backslash \{x\} & \text{for } (x, y) \in E. \end{cases}$$

This shows that

$$\text{(A1.6)} \qquad d(x|G_y) = \begin{cases} d(x|G) & \text{for } (x, y) \notin E, \\ d(x|G) + d(y|G) - 2 - |\text{Adj}\,(x|G) \cap \text{Adj}\,(y|G)| & \text{for } (x, y) \in E. \end{cases}$$

Therefore,

$$\text{(A1.7)} \qquad d(x|G_y) = d(y|G_x) \quad \text{for } (x, y) \in E.$$

From (A1.5) and (A1.6) we derive

$$\text{(A1.8)} \qquad d(x|G_y) < d(x|G) \Rightarrow (x, y) \in E \text{ and } \text{Adj}\,(y|G) \backslash \{x\} \subset \text{Adj}\,(x|G) \backslash \{y\} \text{ and } d(x|G_y) = d(x|G) - 1.$$

LEMMA A1.9. *If $A \subset X$ and $x \in X \backslash A$ then*

$$\text{Adj}\,(x|G) \textit{ complete} \Rightarrow \text{Adj}\,(x|G_A) \textit{ complete}.$$

Proof. The proof is obvious.

The following three lemmata deal with paths in elimination graphs; we set $G = (X, E)$ to be a graph and $A \subset X$.

LEMMA A1.10. *For $x, y \in X \backslash A$, $x \neq y$, the following statements are equivalent*:

(i) *In G exists a path ω from x to y*;

(ii) *In G_A exists a path ω' from x to y.*

The proof of A1.10 is straightforward. It allows us to prove a more extensive statement: For every path ω in G (from x to y) there exists a path ω' in G_A (from x to y) with $Z(\omega') = Z(\omega)\backslash A$. Along with this remark, Lemma A1.10 shows the following.

LEMMA A1.11. *Given distinct vertices $x, y_1, \cdots, y_k \in X\backslash A$ ($k \in \mathbb{N}$) along with disjunct paths ω_i, $i = 1, \cdots, k$, from x to y_i. Then there exist disjunct paths ω_i' in G_A, $i = 1, \cdots, k$, from x to y_i with $Z(\omega_i') = Z(\omega_i)\backslash A$. The converse is not true.*

LEMMA A1.12. *Given distinct vertices $x, y_1, \cdots, y_k$ of G along with disjunct paths ω_i, $i = 1, \cdots, k$, from x to y_i. Then,*

(i) *For any $a \in \operatorname{Adj}(x|G)\backslash\{y_1, \cdots, y_k\}$ there exist disjunct paths ω_i' in G_x from a to y_i with $Z(\omega_i') \subset Z(\omega_i)\backslash\{a\}$.*

(ii) *For any y_i and for any*

$$b \in \operatorname{Adj}(y_i|G)\backslash\left(\{y_1, \cdots, y_k, x\} \cup \bigcup_{\substack{j=1\\ j\neq 1}}^{k} Z(\omega_j)\right):$$

There are disjunct paths ω_j' in G_{y_i}, $j = 1, \cdots, k$, from x to y_j for $j \neq i$, respectively, from x to b for $j = i$, with $Z(\omega_j') \subset Z(\omega_j)\backslash\{b\}$.

Proof. (i) Let $a \in \operatorname{Adj}(x|G)\backslash\{y_1, \cdots, y_k\}$. We define other paths ω_i'', $1 \leqq i \leqq k$: If $a \notin Z(\omega_i)$, set $\omega_i'' := \langle a \rangle + \omega_i$ which may be considered as the "extension of ω_i to a." In the other case ($a \in Z(\omega_t)$) ω_i'' denotes that part of ω_i which leads from a to y_i; this occurs for at most one i. Evidently, we have got paths from a to y_i, $1 \leqq i \leqq k$, which satisfy

— $Z(\omega_i'') \cap Z(\omega_j'') \subset \{x\}$ for $i \neq j$;

— $a, y_j \notin Z(\omega_i'')$ for $i \neq j$;

— $Z(\omega_i'')\backslash\{x\} \subset Z(\omega_i)\backslash\{a\}$.

By elimination of x the paths ω_i'' changes into disjunct paths ω_i' of G_x, $1 \leqq i \leqq k$, with $Z(\omega_i') = Z(\omega_i'')\backslash\{x\} \subset Z(\omega_i)\backslash\{a\}$. Therefore, (i) is proved.

(ii) Given any y_i and any $b \in \operatorname{Adj}(y_i|G)\backslash(\cdots)$. Again we consider other paths ω_j'', $1 \leqq j \leqq k$: Set $\omega_j'' := \omega_j$ for $j \neq i$. In the other case we set: If $b \in Z(\omega_i)$ let ω_i'' be that part of ω_i leading from x to b. If $b \notin Z(\omega_i)$, ω_i'' denotes the extension of ω_i to b, i.e., $\omega_i'' := \omega_i + \langle b \rangle$. We get disjunct paths ω_j'' from x to y_j, for $j \neq i$, respectively, from x to b, for $j = i$. Elimination of y_i preserves the paths ω_j'', $j \neq i$, unchanged; ω_i'' is possibly contracted by y_i (if $b \notin Z(\omega_i)$), but it remains a path from x to b. So (ii) is shown. □

Appendix A2. An elimination ordering is dominating if and only if it is $\mathscr{F}$-minimal. Ranking elimination orderings by criterion functions is compatible with the quasi-ordering "dominates" (Definition 1.2) in the following manner:

$$\alpha \text{ dominates } \alpha' \Rightarrow f(D(\alpha|G)) \leqq f(D(\alpha'|G)) \quad \text{for all } f \in \mathscr{F}.$$

Therefore, each elimination ordering which dominates an f-minimal elimination ordering ($f \in \mathscr{F}$) is f-minimal too. Especially, a dominating elimination ordering is $\mathscr{F}$-minimal. To show the converse a special criterion function, defined below, is used. In addition, the following example demonstrates that $\mathscr{F}$ contains very complicated samples.

Example A2.1. Given positive real numbers $g_1, \cdots, g_n$. The function f defined by

$$f(a_1, \cdots, a_n) := \sum_{i=1}^{n} g_i \cdot a_{\pi(i)},$$

where π is a permutation (depending on $a_1, \cdots, a_n$) of the numbers $1, \cdots, n$ satisfying

$$a_{\pi(1)} \leqq a_{\pi(2)} \leqq \cdots \leqq a_{\pi(n)},$$

is a criterion function.

Proof. First we show the following statement:

For all $a_1, \cdots, a_n, b_1, \cdots, b_n \in \mathbb{N}$ with $a_i \leqq b_i$ and for two corresponding permutations π and ν of the numbers $1, \cdots, n$ which satisfy $a_{\pi(1)} \leqq a_{\pi(2)} \leqq \cdots \leqq a_{\pi(n)}$ and $b_{\nu(1)} \leqq b_{\nu(2)} \leqq \cdots \leqq b_{\nu(n)}$ it is true: $a_{\pi(i)} \leqq b_{\nu(i)}$ for $1 \leqq i \leqq n$.

We assume that there is an i_0, $1 \leqq i_0 \leqq n$ with $b_{\nu(i_0)} < a_{\pi(i_0)}$. Thus, $b_{\nu(i)} < a_{\pi(j)}$ for $1 \leqq i \leqq i_0$ and $i_0 \leqq j \leqq n$ $(*)$. Obviously, to each i, $1 \leqq i \leqq i_0$, there exist $j \in \mathbb{N}$ with $\nu(i) = \pi(j)$. Since $b_{\nu(i)} \geqq a_{\pi(j)}$, we get with $(*)$ (for the chosen j): $j < i_0$. Thus

$$\nu(\{k \in \mathbb{N} | k \leqq i_0\}) \subsetneqq \pi(\{k \in \mathbb{N} | k \leqq i_0\})$$

which is a contradiction to the bijectivity of π and ν. From the statement it follows directly that f is well defined, symmetric, monotone and therefore a criterion function. □

PROPOSITION A2.2. *An elimination ordering is $\mathcal{F}$-minimal if and only if it is dominating.*

Proof. The "if" part is obvious. For the "only if" part let α be a $\mathcal{F}$-minimal and α' be any elimination ordering of G. We employ certain samples f_i of the criterion function defined above; set f_i to be that function with $g_i = 1$ and $g_j = 0$ for $i \neq j$, $1 \leqq i \leqq n$. Furthermore we introduce the following abbreviations: $d_i := d(\alpha(i)|G_{\alpha(1,i-1)})$ and $d'_i := d(\alpha'(i)|G_{\alpha'(1,i-1)})$ for $i = 1, \cdots, n$; thus $D(\alpha|G) = (d_1, \cdots, d_n)$ and $D(\alpha'|G) = (d'_1, \cdots, d'_n)$. Moreover, set π and ν to be permutations of the numbers $1, \cdots, n$ with $d_{\pi(1)} \leqq d_{\pi(2)} \leqq \cdots \leqq d_{\pi(n)}$ and $d'_{\nu(1)} \leqq d'_{\nu(2)} \leqq \cdots \leqq d'_{\nu(n)}$. According to the $\mathcal{F}$-minimality of α we get: $d_{\pi(i)} = f_i(D(\alpha|G)) \leqq f_i(D(\alpha'|G)) = d'_{\nu(i)}$ for $1 \leqq i \leqq n$. Thus, α dominates the (given) elimination ordering α' and the proof is complete. □

The proof above shows furthermore that α is $\mathcal{F}$-minimal if and only if α is f_i-minimal for $i = 1, \cdots, n$. Therefore f_i-minimality for only a finite number of (suitable) criterion functions is sufficient to guarantee $\mathcal{F}$-minimality.

Appendix A3. Proof of Theorem 2.6.

LEMMA A3.1. *Given a graph $F = (Y, L)$ whose complementary graph $\bar{F}$ is a forest of bushes. Then for the root w of any bush of $\bar{F}$ and for one of its peaks b:*

(i) $(w, y) \in L$ *for all* $y \in Y \backslash (B(w) \cup \{w\})$, *where $B(w)$ denotes the set of all peaks of the bush with root w.*

(ii) $(b, y) \in L$ *for all* $y \in Y$, $y \neq w$, $y \neq b$.

Proof. The proof is obvious.

For the following five lemmata let $G = (X, E)$ be a graph; x and z are set to be vertices both of type B and adjacent to each other. Elimination of z may destroy the property B of x. This situation is considered in Lemmata A3.3–A3.5.

LEMMA A3.2. *With the assumptions made above it is true*: $d(x|G) = d(z|G)$.

Proof. From Lemma A3.1 and Definition 2.1 we conclude that $d(z|G) \geqq d(x|G)$, where we have to consider the two situations, z is a root, respectively, z is a peak of a bush of $\overline{G(\text{Adj}\,(x|G))}$, separately. By symmetry we get $d(x|G) \leqq d(z|G)$ too. □

LEMMA A3.3. *In addition to the assumptions made above let z be a peak of a bush of* $\overline{G(\text{Adj}\,(x|G))}$. *Then x is* (*also*) *in G_z of type* B.

Proof. Let w be the root belonging to z. Then: $(*)$ There exists one and only one $v \in \text{Adj}\,(z|G)$, $v \neq x$, with $v \notin \text{Adj}\,(x|G)$. To verify $(*)$ we set

$$A := \text{Adj}\,(z|G) \backslash (\text{Adj}\,(x|G) \cup \{x\})$$

and prove $|A| = 1$. Lemma A3.1 (ii) shows Adj $(x|G)\backslash\{w, z\} \subset$ Adj $(z|G)\backslash(A \cup \{x\})$ and therefore $d(x|G) - 2 \leqq d(z|G) - |A| - 1$. With Lemma A3.2 we derive $|A| \leqq 1$. If $A = \emptyset$ we get Adj $(z|G)\backslash\{x\} \subset$ Adj $(x|G)\backslash\{z, w\}$, respectively, $d(z|G) - 1 \leqq d(x|G) - 2$, which is a contradiction to Lemma A3.2. Summarizing we have proved that $|A| = 1$. Subsequently, it is shown that $\overline{G_z(\text{Adj}\,(x|G_z))}$ is a forest of bushes. To prove this we employ the map Γ : Adj $(x|G) \rightarrow$ Adj $(x|G_z)$ defined by $\Gamma(z) := v$ and $\Gamma(u) := u$ for $u \neq z$. Since Adj $(x|G_z) =$ (Adj $(x|G)\backslash\{z\}) \cup \{v\}$, Γ is surjective; $v \notin$ Adj $(x|G)$ shows Γ injective. Furthermore, $(a, b) \in E \Rightarrow (\Gamma(a), \Gamma(b)) \in E_z$ for all $a, b \in$ Adj $(x|G)$, $a \neq b$. To verify this statement, let $(a, b) \in E$ and (without loss of generality) set $b = z$. Since $(a, z) \in E$ and $(z, v) \in E$ we get: $(a, v) = (\Gamma(a), \Gamma(z)) \in E_z$. Therefore, Γ is an isomorphism between G(Adj $(x|G)$) and a subgraph of G_z(Adj $(x|G_z)$). Consequently, $\overline{G_z(\text{Adj}\,(x|G_z))}$ is (more than ever) a forest of bushes, where v is either a peak belonging to the root w or a root of a trivial bush. Note, that v cannot be a root, respectively, a peak, of any other bush (one with a root unequal to w). Considering in addition Lemma A1.12 (ii), we have proved that x is in G_z of type B. □

LEMMA A3.4. *In addition to the assumptions made above let z be the root of a trivial bush of* $\overline{G(\text{Adj}\,(x|G))}$. *Then x is in G_z of type* B.

Proof. From Lemmata A3.1 and A3.2 we get: Adj $(x|G)\backslash\{z\} =$ Adj $(z|G)\backslash\{x\}$. Since Adj $(x|G_z) =$ Adj $(z|G)\backslash\{x\}$ is complete in G_z, x is of type B in G_z. Therefore, Lemma A3.1 is proved. □

LEMMA A3.5. *In addition to the assumptions made above let z be the root of a bush of* $\overline{G(\text{Adj}\,(x|G))}$ *with at least two peaks. Then, the map* $\Gamma : X\backslash\{x\} \rightarrow X\backslash\{z\}$ *defined by* $\Gamma(z) := x$ *and* $\Gamma(u) := u$ *for* $u \neq z$ *is an isomorphism between the graphs* G_x *and* G_z.

Proof. Set $b_1, \cdots, b_r$, to be the peaks belonging to z $(r \geqq 2)$. Furthermore set $A :=$ Adj $(z|G)\backslash$(Adj $(x|G) \cup \{x\})$. Then

$$\text{Adj}\,(z|G)\backslash(\{x\} \cup A) \subset \text{Adj}\,(x|G)\backslash\{z, b_1, \cdots, b_r\}.$$

From Lemma A3.2 it follows that $|A| \geqq r$. Therefore, x is the root of a bush of $\overline{G(\text{Adj}\,(z|G))}$ with at least two peaks $(*)$.

Now we prove:

$$(u, v) \in E_x \Leftrightarrow (\Gamma(u), \Gamma(v)) \in E_z \quad \text{for all } u, v \in X\backslash\{x\}, u \neq v.$$

"$\Rightarrow$" Let $(u, v) \in E_x$. If $(u, v) \in E$ and $u = z$ (the case $u, v \neq z$ is trivial) then: $(x, v) = (\Gamma(z), \Gamma(v)) \in E_z$ (note: $(z, x) \in E$). If $(u, v) \notin E$ then $(u, x) \in E$ and $(v, x) \in E$. Since x is of type B we set (without loss of generality) u to be the root and v to be a corresponding peak of a bush of $\overline{G(\text{Adj}\,(x|G))}$, especially $v \neq z$. We distinguish between $u \neq z$, respectively, $u = z$. If $u \neq z$ then Lemma A3.1 shows: $(u, z) \in E$ and $(v, z) \in E$. Therefore $(u, v) = (\Gamma(u), \Gamma(v)) \in E_z$. If $u = z$ we conclude with $(v, x) \in E$ that $(\Gamma(u), \Gamma(v)) = (x, v) \in E_z$.

"$\Leftarrow$" Follows directly from "$\Rightarrow$" by considering the inherent symmetry and $(*)$.

Lemmata A3.3, A3.4 and A3.5 take all possible situations for $z \in$ Adj $(x|G)$ into consideration. The case, z is the root of a bush with exactly one peak, is comprised in Lemma A3.3. We summarize in Lemma A3.6.

LEMMA A3.6. *With the assumptions made above, either x is in G_z of type* B *or G_x is isomorphic to G_z.*

With the aid of Lemma A3.6 the statement of Theorem 2.6 can be reduced to a statement onto partial ordered sets; the notation which is used is taken from [Bi67]. We set

$$\mathcal{M} := \{G_\gamma | \gamma \text{ is a B-partial elimination ordering of } G\}.$$

Since the empty elimination ordering can be considered as a B-partial elimination ordering, G is contained in $\mathscr{M}$. Elimination graphs which are isomorphic to each other ($\simeq$) are identified by an equivalence relation $\sim$ on $\mathscr{M}$:

$$F \sim H :\Leftrightarrow F \simeq H (F, H \in \mathscr{M}).$$

The classes of this equivalence relation are denoted by $[F]$, $F \in \mathscr{M}$, the totality of these classes by $\mathscr{K}$. Onto $\mathscr{K}$ a partial ordering is defined by

$$[F] \leqq [H] :\Leftrightarrow \text{There is a B-partial elimination ordering } \gamma \text{ of } H \text{ with } F \simeq H_\gamma.$$

Any B-partial elimination ordering β of G induces a chain $[G] \geqq [G_{\beta(1)}] \geqq [G_{\beta(1,2)}] \geqq \cdots \geqq [G_\beta]$. Especially, $[G_\beta]$ is a minimal element in $\mathscr{K}$ if and only if β is not continuable. Thus, to prove Theorem 2.6 (ii) we have to show that $\mathscr{K}$ contains one and only one minimal element. But this follows directly from Lemma A3.7 because the following condition holds for the partial ordered set $(\mathscr{K}, \leqq)$:

For any two $[F]$, $[H] \in \mathscr{K}$, $[F] \neq [H]$ which are covered by any class there exist a class which is covered by $[F]$ and $[H]$.

This is easily verified: Let $[K]$ be the class covering $[F]$ and $[H]$. Then there are two distinct vertices x, z of K, both of type B, so that (without loss of generality) $F = K_x$ and $H = K_z$. Since F is not isomorphic to H, z is in F of type B and x is in H of type B (A3.6). Therefore $[K_{\langle x,z \rangle}]$ is covered by $[F]$ and $[H]$. Altogether Theorem 2.6 (ii) is proved, and 2.6 (i) is a simple consequence of (ii). □

The following lemma remains to be proved.

LEMMA A3.7. *A finite partial ordered set* $(M, \leqq)$ *satisfying the following conditions contains one and only one minimal element.*

(i) *There exists a largest element* $a \in M$.

(ii) *To any two* $x, y \in M$, $x \neq y$, *which are covered by any* $u \in M$ *there exists* $v \in M$ *which is covered by* x *and* y. (x *covers* y *if and only if* $x \geqq y$ *and there is no other* $z \in M$, $z \neq x$, $z \neq y$, *satisfying* $x \leqq z \leqq y$.)

Proof. The existence of a minimal element is obvious. Furthermore we need the following statement, which may be proved by induction over l where (ii) is to apply:

Let $c = x_0 \geqq x_1 \geqq \cdots \geqq x_l = b$ be a maximal chain from an element c to a minimal element b. Furthermore given some c' which is covered by c. Then there exists a maximal chain $c' = y_0 \geqq y_1 \geqq \cdots \geqq y_{\ell - 1} = b$ from c' to b.

Set b, b' to be minimal elements of M. Since $a \geqq b$ and $a \geqq b'$ there are maximal chains $a = x_0 \geqq x_1 \geqq \cdots \geqq x_\ell = b$ and $a = x'_0 \geqq x'_1 \geqq \cdots \geqq x'_{\ell'} = b'$. Set k to be the maximal subscript satisfying $x_k = x'_s$ for some s, $0 \leqq s \leqq \ell'$. Without loss of generality we assume for the chain from a to b: For all (other) maximal chains $a = z_0 \geqq z_1 \geqq \cdots \geqq z_r = b$ from a to b holds: If $z_i = x'_j$ for any j, $0 \leqq j \leqq \ell'$ then $i \leqq k$ (*). According to the statement made above we see: If $k < \ell$ then there is a maximal chain $x'_{s+1} = y_{k+1} \geqq y_{k+2} \geqq \cdots \geqq y_\ell = b$ from x'_{s+1} to b. Thus $a = x_0 \geqq x_1 \geqq \cdots \geqq x_k \geqq y_{k+1} \geqq \cdots \geqq y_\ell = b$ is a maximal chain from a to b. But this is a contradiction to (*). Therefore, only $k = \ell$, respectively, $b = b'$ is possible.

Appendix A4. Proof of Theorem 4.4. First some lemmata necessary for the proof of Theorem 4.4 are introduced.

LEMMA A4.1. *Let* $G = (X, E)$ *be a graph,* $Y \subset X$, $S \subset X$, $S \cap Y = \varnothing$, $S_0 \subset S$ *complete and* (S, S_0) *of type* τ_k *relative to* Y (*notice*: *Definition* 4.2 (iii) *is necessary only*). *Set* $\operatorname{Adj}_Y(T|G) := \operatorname{Adj}(T|G) \cap Y$. *Then for each set of vertices* $T \subset S \backslash S_0$ *with* $|T| \leqq k$ *and* $|T| \leqq |Y|$,

$$|\operatorname{Adj}_Y(T|G)| \geqq |T|.$$

Proof (without loss of generality set $T \neq \emptyset$). We assume $|\mathrm{Adj}_Y(T|G)| < |T| =: q$ (*). Since $|T| \leqq |Y|$ there exists $x \in Y$ with $x \notin \mathrm{Adj}_Y(T|G)$. According to Definition 4.2 (iii) there are q nontrivial paths ω_t, $t \in T$, from x to t with $Z(\omega_t) \subset Y$. Therefore, $|\mathrm{Adj}_Y(T|G)| \geqq q$ which contradicts (*). □

We remark that the notation $\mathrm{Adj}_Y(T|G)$ is used for any set Y of vertices of G. Furthermore we introduce $d_Y(x|G) := |\mathrm{Adj}_Y(x|G)|$.

LEMMA A4.2. *Let $G = (X, E)$ be a graph, $Y \subset X$, $S \subset X$, $S \cap Y = \emptyset$, $S_0 \subset S$ complete and (S, S_0) of type τ_k relative to Y. Furthermore let $T \subset S\backslash S_0$ with $|Y| \leqq |T| =: q \leqq k$. Then the following statements are true*:

(i) *To each vertex $y \in Y$ there exists $s \in T$ with $(y, s) \in E$.*

(ii) *For any two distinct vertices $s, s' \in S$,*

$$(s,s') \notin E \Rightarrow (\textit{For all } y \in Y\colon (s,y) \in E \wedge (s',y) \in E) \wedge (|Y| = k).$$

(iii) *For any two distinct vertices $y, y' \in Y$ there exist $s, s' \in T$ with*

$$((y,s) \in E \wedge (s,y') \in E) \vee ((y,s) \in E \wedge (s,s') \in E \wedge (s',y') \in E).$$

(iv) *For any $y \in Y$ and any $s \in S$ it is true*

$$(y,s) \notin E \Rightarrow \textit{There exists } s' \in T \textit{ with: } (y,s') \in E \wedge (s',s) \in E.$$

If $|Y| < k$ then for G_T,

(v) Adj *$(y|G_T)$ is complete (in G_T) for all $y \in Y$.*

(vi) *To any $f \in \mathscr{F}(G_T)$ and any set of vertices $V \subset X\backslash(Y \cup T)$ which is complete in G_T there exists an f-minimal elimination ordering α of G_T, which eliminates Y at the beginning and V at the end, i.e., $\alpha(i) \in Y$ for $i = 1, \cdots, |Y|$ and $\alpha(i) \in V$ for $i = |X_T| - |V| + 1, \cdots, |X_T|$.*

Proof. (i) We assume that there exists $y \in Y$ with $(y, s) \notin E$ for all $s \in T$. Then according to Definition 4.2 (iii) there are q disjunct and nontrivial paths ω_t, $t \in T$, from y to t with $Z(\omega_t) \subset Y\backslash\{y\}$. Therefore, $|Y\backslash\{y\}| \geqq q$. But this is a contradiction to $|Y| \leqq q$.

(ii) If $(s, s') \notin E$ then Definition 4.2 (ii) guarantees that there exist k disjunct and nontrivial paths ω_i, $i = 1, \cdots, k$, from s to s' with $Z(\omega_i) \subset Y$. Thus, $|Y| \geqq k$. Considering $|Y| \leqq k$ we get $|Y| = k$. Hence, to every path ω_i there exists one and only one $y_i \in Y$ with $\omega_i = \langle s, y_i, s' \rangle$. This proves (ii).

(iii) According to (i) there are $s, s' \in T$ with $(y, s) \in E$ and $(y', s') \in E$. If $s = s' \vee (y, s') \in E \vee (y', s) \in E$, there is nothing to prove. Now we consider the converse case: $s \neq s' \wedge (y, s') \notin E \wedge (y', s) \notin E$. According to (ii) this is possible only if $(s, s') \in E$. Therefore, (iii) is proved.

(iv) This follows directly from (i) and (ii).

(v) Given any $y \in Y$, from (ii) it follows ($|Y| < k$) that S is complete in G(*). Therefore, Adj $(T|G)$ is complete in G_T (i.e., G_T(Adj $(T|G)$) complete) (**). With (A1.4) and (i) we get: Adj $(y|G_T) =$ Adj $(y|G)\backslash T \cup$ Adj $(T|G)\backslash\{y\}$. With Definition 4.2 (i), (ii) and (*) we conclude: Adj $(y|G)\backslash T \subset (Y \cup S)\backslash(T \cup \{y\}) \subset$ Adj $(T|G)\backslash\{y\}$. Since Adj $(y|G_T) \subset$ Adj $(T|G)$, (**) guarantees that Adj $(y|G_T)$ is complete (in G_T).

(vi) This follows directly from (v), where Theorems 1.4 (Final Theorem) and 1.3 are employed. □

LEMMA A4.3. *Let $G = (X, E)$ be a graph, $Y \subset X$, $S \subset X$, $Y \cap S = \emptyset$, $S_0 \subset S$ complete and (S, S_0) of type τ_k relative to Y. Furthermore set $|Y| \leqq k$, $T \subset S\backslash S_0$ with $|T| = |Y|$, $\hat{Y} \subset Y$, $\hat{T} \subset T$ with $|\hat{T}| = |\hat{Y}|$, $y \in Y\backslash\hat{Y}$ and $s \in T\backslash \hat{T}$. Then the following inequality holds*:

$$d(s|G_{Y \cup \hat{T}}) \leqq d(y|G_{T \cup \hat{Y}}).$$

Proof. First we show:

$$(*) \qquad B := \mathrm{Adj}\,(s|G_{Y\cup\hat{T}})\backslash(\tilde{T}\cup(S\backslash T))\subset\mathrm{Adj}\,(y|G_{T\cup\hat{Y}})\backslash(\tilde{Y}\cup(S\backslash T))$$

where $\tilde{T} := T\backslash(\hat{T}\cup\{s\})$ and $\tilde{Y} := Y\backslash(\hat{Y}\cup\{y\})$. Given any $x \in B$, especially $x \notin \mathrm{Y}$ and $x \notin \mathrm{S}$, using Lemma A4.2 (ii) we see that S is in G_Y complete; thus $s \in \mathrm{Adj}\,(T|G_Y)$. Therefore, with (A1.4) we get $x \in \mathrm{Adj}\,(s|G_Y)\backslash\hat{T}$ or $x \in \mathrm{Adj}\,(\hat{T}|G_Y)\backslash\{s\}$. Consequently $x \in \mathrm{Adj}\,(t_0|G_Y)$ for any $t_0 \in \hat{T}\cup\{s\}$. Applying (A1.4) again we get $x \in \mathrm{Adj}\,(t_0|G)\backslash Y$ or $x \in \mathrm{Adj}\,(Y|G)\backslash\{t_0\}$. Since $x \notin S$ and $\mathrm{Adj}\,(Y|G)\subset S$ only the case $x\in\mathrm{Adj}\,(t_0|G)\backslash Y$ is possible. Lemma A4.2 (iv) guarantees that $(y, t_0) \in E$ or $(y, t_0) \in E_t$ for any $t \in T$. Therefore, we get $(y, x) \in E_T$, respectively, $(y, x) \in E_{T\cup\hat{Y}}$ which proves $(*)$. Since S is complete in G_Y we see: $\tilde{T}\cup(S\backslash T)\subset\mathrm{Adj}\,(s|G_{Y\cup\hat{T}})$. Lemma A4.2 (iii) and (iv) guarantee $\tilde{Y}\cup(S\backslash T)\subset\mathrm{Adj}\,(y|G_{T\cup\hat{Y}})$. Since in addition $|\tilde{T}\cup(S\backslash T)| = |\tilde{Y}\cup(S\backslash T)|$, the statement of Lemma A4.3 follows from $(*)$. □

LEMMA A4.4. *Let $G = (X, E)$ be a graph, $S\subset X$, $S_0\subset S$ complete and (S, S_0) of type T_k. G is split by S in $G(X\backslash S) = G(Y)\oplus G(Z)$. Furthermore set $x \in X\backslash S$, $T\subset S\backslash S_0$, $|Y|$, $|Z| \geqq |T|$ and $|T| \leqq k$. Then*

$$d(x|G)\leqq d(x|G_T).$$

Proof. Without loss of generality let $x \in Y$. Furthermore let T_i, $i = 1, \cdots, l$, be the sets of vertices of that connected components of $G(T)$ which satisfy $x \in \mathrm{Adj}\,(T_i|G)$; set $\hat{T} := \bigcup_{i=1}^{l} T_i$ and $T_0 := \{s \in T|(s, x) \in E\}$. We see $T_0 \subset \hat{T}$. Lemma A4.1 guarantees: $\mathrm{Adj}_Z\,(\hat{T}|G)| \geqq |\hat{T}|$. Finally we conclude with (A1.4):

$$\begin{aligned} d(x|G_T) &= d(x|G)-|T_0|+|\mathrm{Adj}\,(\hat{T}|G)\backslash(\mathrm{Adj}\,(x|G)\cup\{x\})| \\ &\geqq d(x|G)-|T_0|+|\mathrm{Adj}_Z(\hat{T}|G)| \\ &\geqq d(x|G)-|T_0|+|\hat{T}|\geqq d(x|G). \end{aligned}$$

□

LEMMA A4.5. *Let $G = (X, E)$ be a graph, $S\subset X$, $S_0\subset S$ complete, (S, S_0) of type T_k and $x\in X\backslash S$. Then (S, S_0) is (respectively, remains) in G_x of type T_k.*

Proof. The proof is straightforward by Remark 2 on Definition 4.1, Lemma A1.11 and Lemma A4.6.

LEMMA A4.6. *We have a graph $G = (X, E)$ and two distinct vertices a, b of G with $(a, b) \notin E$. Furthermore, we have k disjunct paths ω_i, $i = 1, \cdots, k$, from a to b. Then for any $x \in X$, $x \neq a$, b, the following condition holds:*

$$\begin{aligned}(a,b)\notin E\Rightarrow &\textit{ There exist } k \textit{ disjunct paths } \omega_i', i=1,\cdots,k, \text{ in } G_x, \textit{ from } a \textit{ to } b \\ &\textit{ with } Z(\omega_i')=Z(\omega_i)\backslash\{x\}.\end{aligned}$$

Proof. The proof is obvious.

For the rest of this proof we introduce the following notation.

Notation. To $f\in\mathscr{F}$ set $\mathfrak{A}_f$ to be the set of all f-minimal elimination orderings α of G which satisfy $S_0\subset M(\alpha(n-j+1, n))\subset S$ for at least one $j\in\mathbb{N}$, $1\leqq j\leqq n = |X|$. For each $\alpha\in\mathfrak{A}_f$ the number

$$l(\alpha) := \max\{j\in\mathbb{N}|S_0\subset M(\alpha(n-j+1,n))\subset S\}$$

is well defined. Further we write down two trivial identities which are used in the following:

$$|\alpha(n-\ell+1,n)_S| = \ell \quad \text{for } \ell := l(\alpha) \quad \text{and}$$

$$|\alpha(1,j)_S|+|\alpha(j+1,n)_S| = |S| = m \quad \text{for } j=1,\cdots,n.$$

Now we are able to prove the statement of Theorem 4.4.

Proof. Without loss of generality we assume $k > 0$, $Y \neq \emptyset$, $Z \neq \emptyset$. Let $f \in \mathcal{F}$ and $\mathfrak{A} = \mathfrak{A}_f$. Theorem 1.4 (Final Theorem) guarantees $\mathfrak{A} \neq \emptyset$. If $l(\alpha') = m = |S|$ for at least one $\alpha' \in \mathfrak{A}$ the statement of Theorem 4.4 follows directly from Theorem 1.4. Therefore, we assume for the rest of this proof that

$$l(\alpha') < m \quad \text{for all } \alpha' \in \mathfrak{A}, \tag{1}$$

which will result in a contradiction. Since $\mathfrak{A} \neq \emptyset$ there exists $\alpha \in \mathfrak{A}$ satisfying: $l(\alpha) \geqq l(\alpha')$ for all $\alpha' \in \mathfrak{A}$. For (that) α,

$$\ell := l(\alpha) \geqq m - k. \tag{2}$$

To prove (2) we assume that $\ell < m - k$, i.e., $|\alpha(1, n - \ell)_S| > k$. Thus, there exists $j \in \mathbb{N}$, $1 \leqq j < n - \ell$, satisfying $|\alpha(1, j)_S| = k$ and $\alpha(j) \in S$. Set $\sigma := \alpha(1, j)_S$ and $G' := G_{\alpha(1,j) - \sigma}$. For σ, $M(\sigma) \subset M(\alpha(1, n - l)_S) = S \backslash M(\alpha(n - \ell + 1, n)) \subset S \backslash S_0$. Since S is in G (and especially in G') reducible to a complete graph by elimination of any k vertices out of $S \backslash S_0$ we see that $S \backslash M(\sigma)$ is complete in $G'_\sigma = G_{\alpha(1,j)}$. According to Theorem 1.4 there exists an f-minimal elimination ordering α' of G which eliminates $S \backslash M(\sigma)$ at the end. Since $S_0 \subset S \backslash M(\sigma)$ we get $\alpha' \in \mathfrak{A}$, where $l(\alpha') = |S \backslash M(\sigma)| = m - k > \ell$. But this is a contradiction to the minimality of ℓ. Therefore, (2) is proved.

Now, for the elimination ordering α chosen above, we set p to be the maximal number so that one of the following two conditions holds:

$$|\alpha(p+1, n)_Y| = |\alpha(1,p)_S| \wedge |\alpha(p+1, n)_Z| > |\alpha(1,p)_S|, \tag{A}$$

$$|\alpha(p+1, n)_Z| = |\alpha(1,p)_S| \wedge |\alpha(p+1, n)_Y| \geqq |\alpha(1,p)_S|. \tag{B}$$

In order to prove that $p \in \mathbb{N}$, $1 \leqq p \leqq n - \ell$, exists we consider the three maps $i \mapsto |\alpha(i + 1, n)_Y|$, $i \mapsto |\alpha(i + 1, n)_Z|$ and $i \mapsto |\alpha(1, i)_S|$. The first and the second are monotonely decreasing, the third is monotonely increasing. Their values at $i = 0$, respectively, $i = n - \ell$, are: $|\alpha(1, n)_Y| = |Y| > 0$, $|\alpha(n - \ell + 1, n)_Y| = |\langle \ \rangle| = 0$, $|\alpha(1, n)_Z| = |Z| > 0$, $|\alpha(n - \ell + 1, n)_Z| = |\langle \ \rangle| = 0$, $|\alpha(1, 0)_S| = 0$ and $|\alpha(1, n - \ell)_S| = m - \ell > 0$. Since $X = Y \dot{\cup} S \dot{\cup} Z$, the transition from i to $i + 1$ causes exactly one of these maps to change its value (increase, respectively, decrease) at an amount of one. This guarantees that there exists $p \in \mathbb{N}$ which satisfies B or the following condition:

$$|\alpha(p+1, n)_Y| = |\alpha(1,p)_S| \wedge |\alpha(p+1, n)_Z| \geqq |\alpha(1,p)_S|. \tag{A$_0$}$$

Since $A_0 \vee B$ is true (for p) if and only if $A \vee B$ is true (for p), we have proved that there exists $p \in \mathbb{N}$, $1 \leqq p \leqq n - \ell$, satisfying one of the conditions A or B. Especially, there exists p which is maximal.

Set $q_0 := |\alpha(1, p)_S|$. Since $q_0 \leqq |\alpha(1, n - \ell)_S| = m - \ell$ we derive from (2):

$$q_0 \leqq k. \tag{3}$$

Since $l(\alpha)$ is maximal the inequality $q_0 \leqq m - \ell$ yields an upper bound for the map $l(\ \ )$ defined on $\mathfrak{A}$:

$$l(\alpha') \leqq m - q_0 \quad \text{for all } \alpha' \in \mathfrak{A}. \tag{4}$$

Now we state: There exists a f-minimal elimination ordering α' of G which satisfies one of the following two conditions:

$$\begin{aligned}
&|\alpha'(1,p)_S| = q_0 = |\alpha'(p+1, n)_Y|, \\
&\alpha'(i) \in Y \text{ for } i = p+1, \cdots, p + q_0, \\
&|\alpha'(i, n)_Z| \geqq q_0 \text{ for } i = 1, \cdots, p + q_0, \\
&\alpha' \text{ eliminates } S_0 \text{ at the end,}
\end{aligned} \tag{5$'$}$$

$$(5'')\qquad \begin{aligned}&|\alpha'(1,p)_S| = q_0 = |\alpha'(p+1,n)_Z|,\\ &\alpha'(i)\in Z \text{ for } i = p+1, \cdots, p+q_0,\\ &|\alpha'(i,n)_Y| \geqq q_0 \text{ for } i = 1, \cdots, p+q_0,\\ &\alpha' \text{ eliminates } S_0 \text{ at the end.}\end{aligned}$$

We remark that Y, respectively, Z, are entirely eliminated "from position $p + q_0 + 1$." To prove the statement above we consider the two cases that conditions A and B hold, respectively.

Case A. p satisfies condition A.

Set $\sigma := \alpha(1, p)_S$, $G_{\alpha(1,p)-\sigma} =: G' = (X', E')$, $Y' := Y \cap X'$ and $f' := f_{D(\alpha(1,p)|G)}$. According to Lemma A4.5, (S, S_0) is in G' of type τ_k relative to Y'. Furthermore, $|Y'| = q_0$. We consider the two cases $q_0 < k$, respectively, $q_0 = k$, separately: the case $q_0 > k$ is excluded by (3). If $q_0 < k$, then Lemma A4.2 (vi) guarantees that there exists a f'-minimal elimination ordering γ of $G'_\sigma = G_{\alpha(1,p)}$, which eliminates S_0 at the end and Y' at the beginning. Therefore, the elimination ordering $\alpha' := \alpha(1,p) + \gamma$ is f-minimal and satisfies condition (5'). Now let $q_0 = k$, hence $\ell \leqq m - k$. Considering (2) we get $\ell = m - k$ and therefore $|\alpha(p + 1, n - \ell)_S| = 0$. From $|\alpha(p + 1, n)_Y| = k > 0$ and $|\alpha(n - \ell + 1, n)_Y| = 0$ we have: $M(\alpha(p + 1, n - \ell)) \neq \emptyset$. Altogether, for $\alpha(p + 1)$ only the two cases $\alpha(p + 1) \in Y$ and $\alpha(p + 1) \in Z$ are possible. If $\alpha(p + 1) \in Z$, we derive from condition A and from the fact that p is (especially) maximal relative to condition A: $|\alpha(p + 2, n)_Z| = k$. Therefore $p + 1$ satisfies condition B. But this is a contradiction to the maximality of p. Thus, $\alpha(p + 1) \in Y$ is possible only. Analogously to the case $q_0 < k$ there exists an $f_{D(\alpha(1,p+1)|G)}$-minimal elimination ordering γ of $G_{\alpha(1,p+1)}$ which eliminates S_0 at the end and $Y'\backslash\{\alpha(p + 1)\}$ at the beginning. Consequently, the elimination ordering $\alpha' := \alpha(1, p + 1) + \gamma$ is f-minimal and satisfies (5').

Case B. p satisfies condition B, can be treated analogously because conditions A and B are (nearly) symmetric.

Now we set α to be any elimination ordering satisfying one of the conditions (5'), respectively, (5''); we abbreviate $q := q_0$. According to the statement proved above such α exists. We remark that α should not be mixed up with the elimination ordering introduced at the beginning of this proof. In the following we show that α is dominated by the elimination ordering

$$\beta := \gamma + \alpha(p+1, p+q) + \sigma + \alpha(p+q+1, n),$$

where

$$\sigma := \alpha(1,p)_S \quad\text{and}\quad \gamma := \alpha(1,p) - \sigma.$$

According to the symmetry of (5') and (5'') it is sufficient to consider only the case that α satisfies (5'). If $q = 0$ there is nothing to prove. Therefore, let $q > 0$. In order to simplify the presentation we abbreviate $\sigma = \langle s_1, \cdots, s_q\rangle$, where $s_j := \alpha(r_j)$ and $1 \leqq r_1 < r_2 < \cdots < r_q \leqq p$; for $\alpha(p + 1, p + q)$ we set: $\alpha(p + 1, p + q) = \langle y_1, \cdots, y_q\rangle$ with $y_j := \alpha(p + j)$, $j = 1, \cdots, q$. Additionally, remember that $\alpha \in \mathfrak{A}$ and $M(\sigma) \subset S\backslash S_0$. We begin the proof that β dominates α with

(6) For any $s_j = \alpha(r_j)$, $1 \leqq j \leqq q - 1$, there exists $s \in M(\alpha(r_j + 1, n)_S)$ with $(s_j, s) \notin E_{\alpha(1,r_j-1)}$.

To verify (6) we assume: There exists at least one s_j, $1 \leqq j \leqq q - 1$ with: $(s_j, s) \in E_{\alpha(1,r_j-1)}$ for all $s \in M(\alpha(r_j + 1, n)_S)$. Consequently, the set of vertices $M(\alpha(r_j + 1, n)_S)$ is complete in $G_{\alpha(1,r_j)}$. Therefore, there exists an f-minimal elimination ordering α' of G which eliminates $M(\alpha(r_j + 1, n)_S)$ at the end. Since $S_0 \subset M(\alpha(r_j + 1, n)_S)$ we see $\alpha' \in \mathfrak{A}$.

For $l(\alpha')$ we get: $l = (\alpha') \geqq |M(\alpha(r_j + 1, n)_S)| = m - |\alpha(1, r_j)_S| = m - j \geqq m - q + 1 > m - q_0$. Yet this is a contradiction to (4).

Subsequently, we consider the degrees of the vertices s_j in the corresponding elimination graph $G_{\alpha(1, r_j - 1)}$:

$$d(s_j | G_{\alpha(1,r_j-1)}) \geqq m + k - j \quad \text{for } j = 1, \cdots, q - 1, \tag{7'}$$

$$d(s_q | G_{\alpha(1,r_q-1)}) \geqq m. \tag{7''}$$

To verify (7′) set $\sigma' := \alpha(1, r_j - 1)_S$ for any (fixed) $j \leqq q - 1$, $G_{\alpha(1,r_j-1)-\sigma'} =: G' = (X', E')$, $Y' := X' \cap Y$, $Z' := X' \cap Z$. (S, S_0) is in G' of type τ_k relative to Y' (Lemma A4.5). Condition (6) guarantees: $(s_j, s) \notin E'$ for at least one $s \in M(\alpha(r_j + 1, n)_S)$. Therefore, we conclude with Definition 3.2 (ii): $d_Y(s_j | G') \geqq k$. Definition 4.3 (v) guarantees $d_{S \cup Z'}(s_j | G') \geqq m - 1$. Altogether, we get $d(s_j | G') \geqq m - 1 + k$ because $Y' \cap (S \cup Z') = \emptyset$. Now we set $T := \{s_i \in M(\sigma') | (s_i, s_j) \in E'\}$, $|T| =: t$. In $G'_{\sigma'} = G_{\alpha(1,r_j-1)}$ it is true:

$$d(s_j | G_{\alpha(1,r_j-1)}) \geqq m - 1 + k - t.$$

Since $t \leqq |\sigma'| = j - 1$, the proof of (7′) is complete. For the case $j = q$ we set G', Y', Z', T and t as above. Lemma A4.1 shows $d_{Y'}(T \cup \{s_q\} | G') \geqq t + 1$, condition 4.3 (v) guarantees $d_{S \cup Z'}(s_q | G') \geqq m' - 1$. Analogously to the case $j < q$ we derive $d(s_q | G_{\alpha(1,r_q-1)}) \geqq m - 1 + t + 1 - t = m$. Therefore, (7″) is proved too.

Applying (7′) and (7″) we get a first comparison between α and β.

$$d(s_j | G_{\alpha(1,r_j-1)}) \geqq d(y_j | G_{\gamma + \alpha(p+1,p+j-1)}) \quad \text{for } 1 \leqq j \leqq q. \tag{8}$$

In order to prove (8) we set $G_{\gamma + \alpha(p+1,p+j-1)} =: G' = (X', E')$ for any j, $1 \leqq j \leqq q$, and $Y' = X' \cap Y$. It is sufficient to show: $d(y_j | G') \leqq m + q - j$ $(*)$. Since (in G') Y' is split off by S we see: Adj $(y_j | G') \subset S \cup (Y' \backslash \{y_j\})$. This proves $(*)$ because $|Y' \backslash \{y_j\}| = q - (j - 1) - 1 = q - j$ and $|S| = m$.

A second comparison between α and β follows directly from Lemma A4.4:

$$d(\alpha(i) | G_{\alpha(1,i-1)}) \geqq d(\alpha(i) | G_{\alpha(1,i-1)-\sigma}) \quad \text{for all } \alpha(i) \in M(\alpha(1,p) - \sigma). \tag{9}$$

Another comparison is:

$$d(y_j | G_{\alpha(1,p+j-1)}) \geqq d(s_j | G_{\gamma + \alpha(p+1,p+q) + \sigma(1,j-1)}) \quad \text{for } 1 \leqq j \leqq q. \tag{10}$$

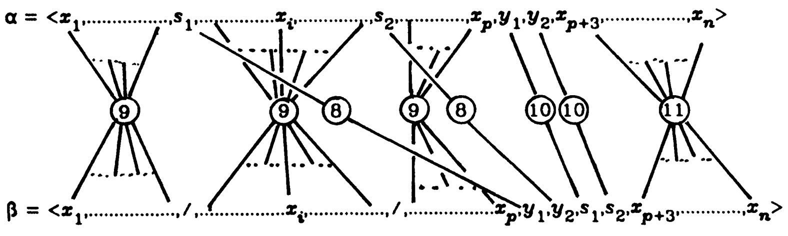
Inequality (10) follows directly from Lemma A4.3 (with $G = G_\gamma$), where we have to consider that $G_{\alpha(1,p+j-1)} = G_{\gamma + \sigma + \alpha(p+1,p+j-1)}$.

Since $M(\alpha(1, p + q)) = M(\gamma + \alpha(p + 1, p + q) + \sigma)$ we see that

$$d(\alpha(i) | G_{\alpha(1,i-1)}) = d(\alpha(i) | G_{\gamma + \alpha(p+1,p+q) + \sigma + \alpha(p+q+1,i-1)}) \qquad \text{for all } i,\ p + q + 1 \leqq i \leqq n. \tag{11}$$

Altogether, from (8)–(11) it follows that β dominates α (Fig. 19 illustrates this for $q = 2$; the corresponding comparisons are denoted by their numbers in the proof). Obviously, β is f-minimal and S is complete in $G_{\gamma + \alpha(p+1,p+q)}$. Consequently, there exists an f-minimal elimination ordering α'' of G which eliminates S at the end. But this is a contradiction to (1). Thus, if condition (1) holds there is no f-minimal elimination ordering $\alpha \in \mathfrak{A}$ satisfying (5′). Analogously it can also be verified that there is no f-minimal elimination ordering $\alpha \in \mathfrak{A}$ satisfying (5″). Altogether Theorem 4.4 is proved. □

Appendix A5. Proof of Theorem 5.7. The following proof of Theorem 5.7 requires some additional notation: To each set A of vertices of G, $A^+ := \{a^* | a \in A\}$ denotes the corresponding set of blocks (of G^*); the set of vertices (of G^*) belonging to A^+ is denoted


α = <x1,...,s1,...,xi,...,s2,...,xp,y1,y2,xp+3,...,xn>
9
9
8
9
8
10
10
11
β = <x1,...,/,...,xi,...,/,...,xp,y1,y2,s1,s2,xp+3,...,xn>

/ indicates erasing of s_j, $j = 1, 2$.

FIG. 19

by $A^* = \bigcup_{a \in A^+} a^*$. Consequently, Adj $(x|G)^+$ is the set of blocks and Adj $(x|G)^*$ is the set of vertices "adjacent" to x^*; especially Adj $(x|G)^* =$ Adj $(x^*|G^*)$.

Adjacency between two blocks a^*, b^* is written in a symbolic manner as $(a^*, b^*) \in E^*$. An extension of an elimination ordering α of G has been defined by

$$\alpha^* = \langle \alpha(1)^* \rangle + \cdots + \langle \alpha(n)^* \rangle.$$

In order to save brackets we denote the orderings $\langle \alpha(j)^* \rangle$ (which are not of particular interest) by $\alpha(j)^*$ too; a confusion between the two meanings of the symbol $\alpha(j)^*$ is not to be expected. Special sections $\alpha(i, j)^*$ are defined by $\alpha(i, j)^* = \alpha(i)^* + \alpha(i+1)^* + \cdots + \alpha(j)^*$. Given a block $x^* = \langle v_1, \cdots, v_c \rangle$ of G^*, $c :=$ dim (x^*). The vector $d(x^*|G^*) := (d(v_i)|G_{\langle v_1, \cdots, v_{i-1} \rangle})_{i=1,\cdots,c}$ combines the degrees of vertices of x^* which arise during the elimination (one by one) of x^*. Remember that $d(x^*|G^*)$ is independent of the employed (internal) ordering. An inequality $d(a^*|G^*) \leqq d(b^*|G^*)$ between two of such vectors is considered to hold for pairs of corresponding entries; note that dim $(a^*) =$ dim (b^*) is necessary.

Proof of Theorem 5.7. Given any elimination ordering α' of G^*. Furthermore set $S :=$ Adj $(x|G)$, $S_0 := \{y \in S|\ y$ is a peak of $\overline{G(\text{Adj}\ (x|G))}\}$, $m := |S|$, $n := |X|$. To α' there exists an elimination ordering α^* satisfying:

α^* is eliminating block by block;

α^* dominates α';

$S_0^* \subset M(\alpha(n-m+2, n)^*) \subset S^*$ (i.e., α^* eliminates $m - 1$ blocks of S^* at the end; especially, S_0^* is eliminated at the end).

This is easily verified by Theorem 2.3. In detail we have to take into consideration that

— S_0^* is a clique in G^*;

— The elimination of a block y^*, where y is a root of

$$\overline{G(\text{Adj}\ (x|G))} \qquad (y^* \in S^+ \backslash S_0^+),$$

causes that the set of vertices $S^* \backslash y^*$ becomes a clique in $G_{y^*}^*$ (Lemma A3.1).

If α^* even satisfies

$$(1) \qquad S_0^* \subset M(\alpha(n-m+1, n)^*) \subset S^*$$

(meaning that α^* eliminates the entire set S^* at the end) then Theorem 1.5 guarantees that there is an elimination ordering which starts with x^* and which is equivalent to α^*, respectively, which dominates α'. Otherwise

— Let y^* be the unique block which is not eliminated at the end of α^*. Note that y^* corresponds to a root of $\overline{G(\text{Adj}\ (x|G))}$:

— k and ℓ are defined by $\alpha(k)^* = y^*$ and $\alpha(\ell)^* = x^*$;

— $x^* = \langle x_1, \cdots, x_c \rangle$ is split into two distinct blocks $u^* := \langle x_1, \cdots, x_s \rangle$ and $v^* := \langle x_{s+1}, \cdots, x_c \rangle$, where $c := \dim(x^*)$ and $s := \dim(y)^*$. Note: Definition 5.6 (ii) guarantees $s \leqq c$.

We have to consider the two cases $\ell < k$ resp. $\ell > k$. If $\ell < k$ (meaning that α^* eliminates x^* before y^*) the elimination of x^* causes S^* to become a clique in $G^*_{\alpha(1,\ell)^*}$. Consequently, there exists an elimination ordering which eliminates S^* at the end and dominates α^*. Therefore, there is an elimination ordering starting with x^* and dominating α'. In the other case ($\ell > k$) we define the elimination ordering

$$\beta := \alpha(1, k-1)^* + u^* + \alpha(k+1, \ell-1)^* + y^* + v^* + \alpha(\ell+1, n)^*$$

and prove subsequently that β dominates α^*. Since $G(\text{Adj}\ (v^* | G^*_{\alpha(1,k-1)^*+u^*}))$ is a clique the elimination ordering $\gamma := \alpha(1, k-1)^* + u^* + v^* + \alpha(k+1, \ell-1)^* + y^* + \alpha(\ell - 1, n)^*$ dominates β (Theorems 2.3 and 1.5). Again, there exists an elimination ordering starting with x^* and dominating α'.

To complete the proof we have to verify that β dominates α^*. We abbreviate: $\hat{G} := G_{\alpha(1,k-1)^*}$. A first comparison between α^* and β is given by

$$d(u^* | \hat{G}) \leqq d(y^* | \hat{G}), \tag{2}$$

which is easily proved by comparing corresponding entries of $d(u^* | G)$ and $d(y^* | G)$: Definition 5.6 (iii) guarantees $d(x_1 | \hat{G}) =: d_1 \leqq d_2 := d(y_1 | \hat{G})$. Consequently, $d(x_2 | \hat{G}_{x_1}) = d_1 - 1 \leqq d_2 - 1 = d(y_1 | \hat{G}_{y_1})$, $d(x_3 | \hat{G}_{\langle x_1, x_2 \rangle}) = d_1 - 2 \leqq d_2 - 2 = d(y_3 | \hat{G}_{\langle y_1, y_2 \rangle})$ and so on. Another comparison between α^* and β, is prepared by the following statements:

Given any set of blocks A^+ with $A^+ \cap (S^+ \cup \{u^*, v^*\}) = \emptyset$. Set $\hat{G}_{A^*} =: H = (Y, L) :=$. Then for any block $z^* \notin A^+ \cup S^+ \cup \{u^*, v^*\}$ it is true:

$$\text{Adj}\ (z^* | H_{u^*}) \backslash y^* \subset \text{Adj}\ (z^* | H_{y^*}) \backslash u^*, \tag{3}$$

$$y^* \subset \text{Adj}\ (z^* | H_{u^*}) \Leftrightarrow u^* \subset \text{Adj}\ (z^* | H_{y^*}), \tag{4}$$

$$d(z^* | H_{u^*}) \leqq d(z^* | H_{y^*}). \tag{5}$$

To prove (3) let $(w^*, z^*) \in L_{u^*}$, $w^* \neq y^*$. If $(w^*, z^*) \in L$ there is nothing to show. In the other case ($(w^*, z^*) \notin L$) we have $(w^*, u^*) \in L$ and $(u^*, z^*) \in L$ which contradicts $z^* \notin S^+$. To verify (4) let $(y^*, z^*) \in L_{u^*}$. If $(y^*, z^*) \in L$ we get $(z^*, u^*) \in L_{y^*}$ because $(y^*, u^*) \in L$. Otherwise ($(y^*, z^*) \notin L$) we have $(y^*, u^*) \in L$ and $(u^*, z^*) \in L$, especially $(u^*, z^*) \in L_{y^*}$. Now let $(u^*, z^*) \in L_{y^*}$. Since $z^* \notin S^+$ we get $(u^*, z^*) \notin L$. Therefore $(u^*, y^*) \in L$ and $(y^*, z^*) \in L$; especially $(y^*, z^*) \in L_{u^*}$. Obviously, (5) follows directly from (3) and (4).

The second comparison between α and β,

$$d(\alpha(j)^* | \hat{G}_{u^* + \alpha(k+1, j-1)^*}) \leqq d(\alpha(j)^* | \hat{G}_{y^* + \alpha(k+1, j-1)^*}), \tag{6}$$

can be derived directly from (5) by setting $\alpha(j)^* = z^*$ and $\hat{G}_{(k+1,j-1)^*} = H$. A third comparison,

$$d(y^* | \hat{G}_{u^* + \alpha(k+1, \ell-1)^*}) = d(u^* | \hat{G}_{y^* + \alpha(k+1, \ell-1)^*}), \tag{7}$$

follows from (A1.7), respectively, from its generalization to simplex graphs. Finally, (2), (6) and (7) guarantee that β dominates α^*. Therefore, Theorem 5.7 is proved. □

REFERENCES

[Bau67] R. BAUMANN, *Topologische Steuerung bei Eliminationsverfahren zur Lastfluberechnung*, Comunicari la simpozionul, Bucharest, 1967.

[Be71] U. BERTELÈ AND F. BRIOSCHI, *On the theory of the elimination process*, J. Math. Anal. Appl., 35 (1971), pp. 48–57.

[Be72] ———, *Nonserial Dynamic Programming*, Academic Press, New York, 1971.

[Bi67] G. BIRKHOFF, *Lattice Theory*, American Mathematical Society, Providence, RI, 1967.

[Ev75] S. EVEN AND E. TARJAN, *Network flow and testing graph connectivity*, SIAM J. Comput., 4 (1975), pp. 507–518.

[Ge73] A. GEORGE, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 539–548.

[Ge78] ———, *An automatic nested dissection algorithm for irregular finite element problems*, SIAM J. Numer. Anal., 15 (1978), pp. 1053–1069.

[Ge80] ———, *An automatic one-way dissection algorithm for irregular finite element problems*, SIAM J. Numer. Anal., 17 (1980), pp. 740–751.

[Ti73] W. F. TINNEY AND W. S. SCOTT, *Solution of large sparse systems by ordered triangular factorization*, IEEE Trans. Automat. Control, 18 (1973), pp. 333–346.

[Pa61] S. PARTER, *The use of linear graphs in Gauss elimination*, SIAM Rev., 3 (1961), pp. 119–130.

[Ro70] D. J. ROSE, *Symmetric elimination on sparse positive definite systems and the potential network flow problem*, Ph.D. thesis, Dept. of Engineering and Applied Physics, Harvard Univ., Cambridge, MA, 1970.

[Ro72] ———, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1973, pp. 183–217.

[Ro76] 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.

[Sch80] H. R. SCHWARZ, *Methode der finiten Elemente*, Teubner, Stuttgart, 1980.

[St74] B. STOTT, *Review of load-flow calculation methods*, Proc. IEEE, 62 (1974).

[We83] H. WENDEL, *Zur Bestimmung optimaler Diagonalpivotfolgen für schwach besetzte, positiv definite Matrizen*, Dissertation, Technische Universität München TUM-I8308, 1983.

[Ya81] M. YANNAKAKIS, *Computing the minimum fill-in is* NP-*complete*, this Journal, 2 (1981), pp. 77–79.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

018

MIXING RATES FOR A RANDOM WALK ON THE CUBE*

PETER MATTHEWS†

Abstract. For a simple random walk on the cube a coupling and a strong uniform time are given. The coupling gives an upper bound on the variation distance between the distribution after k steps and the uniform distribution that is almost the best possible. The strong uniform time is used to calculate the variation distance and the separation. The coupling and strong uniform time are intimately related to a hitting time for the Ehrenfest chain and the time taken by a random graph to become connected, respectively.

Key words. coupling, strong uniform time, random graph, Ehrenfest chain

AMS(MOS) subject classifications. 60B15, 60C05

1. Introduction. The N-cube is the group Z_2^N. It will be convenient to think of the group elements as vectors of zeros and ones of length N. The group operation is then coordinatewise addition modulo two. If $z \in Z_2^N$, $z(i)$ denotes the ith coordinate of z. The random walk $\{X_k, k = 0, 1, \cdots\}$ considered here takes independent steps with distribution μ putting mass

$$(1.1)\qquad \begin{array}{ll} \dfrac{1}{N+1} & \text{on } (0, \cdots, 0), \\[2ex] \dfrac{1}{N+1} & \text{on each of } (1,0,\cdots,0), (0,1,0,\cdots,0), \cdots, (0,\cdots,0,1). \end{array}$$

At each step, either one coordinate changes or nothing happens, all with equal probability. It will frequently be advantageous to think of a fictional $(N+1)$st coordinate that moves each time X does not. This random walk on Z_2^{N+1} will be denoted X^*. Let $m_k(X^*)$ denote the coordinate in which X_{k-1}^* and X_k^* differ.

Starting at $X_0 = (0, \cdots, 0)$, the distribution of the position X_k of the random walk after k steps is the k-fold convolution μ^{k*}. As $k \to \infty$, μ^{k*} converges to the uniform distribution U on Z_2^N. Two measures of the rate of convergence are the variation distance

$$(1.2)\qquad d(\mu^{k*}, U) = \max_{A \subset Z_2^N} |\mu^{k*}(A) - U(A)|$$

and the separation

$$(1.3)\qquad s(\mu^{k*}, U) = \max_{z \in Z_2^N} 2^N\left(\frac{1}{2^N} - \mu^{k*}(z)\right).$$

This example is fairly unusual in that fairly precise calculations of (1.2) and (1.3) can be given. For (1.2) first note that the distribution μ^{k*} is invariant under permutations of the coordinates. Let $|z|$ denote the number of ones in z, for $z \in Z_2^N$. It follows that $\mu^{k*}(z)$ depends on z only through $|z|$, so attention can be restricted to a simpler Markov chain 0, $|X_1|$, $|X_2|$, $\cdots$. This is essentially the classical Ehrenfest chain. Let $K =$

* Received by the editors October 13, 1986; accepted for publication June 8, 1987.

† Department of Mathematics, University of Maryland Baltimore County, Catonsville, Maryland 21228.

$(N/4)(\log N + c)$. Then the classical analysis of the Ehrenfest chain as in Kemperman [6], the more modern Fourier version of the analysis as in Diaconis [4], or a Poissonization argument shows that

$$(1.4) \qquad |X_K| \text{ is approximately binomial } \left(N, \frac{1}{2}\left(1 - \frac{e^{-c/2}}{\sqrt{N}}\right)\right).$$

The number of ones in a uniformly distributed member of Z_2^N is binomial $(N, \frac{1}{2})$. It follows easily that for c fixed, as $N \to \infty$

$$(1.5) \qquad d(\mu^{K*}, U) = 2\Phi(\tfrac{1}{2}e^{-c/2}) - 1 + o(1).$$

For the separation (1.3), the same analyses show that the maximum occurs at $z = (1, \cdots, 1)$. Let $J = (N/2)(\log N + b)$. Then for b fixed, as $N \to \infty$

$$(1.6) \qquad s(\mu^{J*}, u) = 1 - e^{-e^{-b}} + o(1).$$

Both (1.5) and (1.6) exhibit threshold behavior as discussed in Aldous and Diaconis [1]; there are drastic drops in variation distance and separation from about 1 to about 0 near $(N/4) \log N$ and $(N/2) \log N$, respectively. The drop-off points differ by a factor of two, the largest possible [2]. Also, the behaviors as functions of b and c differ; (1.6) drops like an extreme value ($\log \chi_2^2$) tail probability in b, while (1.5) drops like a $\log \chi_1^2$ tail probability in c.

Modern methods of bounding $d(\mu^{k*}, u)$ and $s(\mu^{k*}, u)$ (see [4]) include coupling and strong uniform times. In the present context a coupling is X along with another process Y and a coupling time T such that Y_0 is uniformly distributed, Y has the same transition probabilities as X and $X_k = Y_k$ if $k \geqq T$. For any coupling

$$(1.7) \qquad d(\mu^{k*}, U) \leqq P(T > k)$$

and there is a coupling that attains equality in (1.7) for all k [5]. A strong uniform time is a randomized stopping time T such that

$$P(X_k = z | T = k) = P(X_k = z | T \leqq k) = \frac{1}{2^N} \quad \text{for all } z, k.$$

Strong uniform times are useful because for any strong uniform T

$$(1.8) \qquad s(\mu^{k*}, u) \leqq P(T > k),$$

and there is a strong uniform time attaining equality in (1.8) for all k [2].

As discussed in Aldous and Diaconis [1] a coupling and a strong uniform time from which (1.5) and (1.6), respectively, can be obtained as upper bounds are unknown. These shortcomings place the practical usefulness of these techniques in some doubt. This paper hopes to partially rescue these techniques by giving a coupling and a strong uniform time that give results like (1.5) and (1.6) via (1.7) and (1.8).

Section 2 gives the coupling construction and the result

$$(1.9) \qquad d(\mu^{K*}, U) \leqq 2\Phi\left(\frac{1}{\sqrt{2}} e^{-c/2}\right) - 1 + o(1) \quad \text{as } N \to \infty.$$

This gives the same threshold as (1.5) and an upper bound that is off only by a factor of $\sqrt{2}$ for large c. The coupling is non-Markovian; $\{X, Y\}$ is not jointly a Markov process. Nonetheless, the coupling and the calculation of (1.9) are straightforward. The coupling time is closely related to the time taken by the Ehrenfest chain to reach equilibrium.

The strong uniform time is given in § 3. It is precisely the time taken by a random graph with $N + 1$ vertices, adding an edge randomly at each step with duplicate edges possible, to become connected. Results on random graphs as in Bollobas [3] give the upper bound

$$s(\mu^{J*}, U) \leqq 1 - e^{-e^{-b}} + o(1) \quad \text{as } N \to \infty. \tag{1.10}$$

Separation is typically not of interest in its own right, but rather as an upper bound on variation distance, since $d(\mu^{k*}, U) \leqq s(\mu^{k*}, U)$ [2]. In this example, from (1.5) and (1.6) it is clear that a strong uniform time cannot give a good upper bound on the variation distance. However there is often a lot of structure to a strong uniform time. In many practical examples a strong uniform time T is the last in a sequence of stopping times $T_1, \cdots, T_M$. The distribution of X_k, given that $T_1, \cdots, T_i$ have occurred, has some property that makes it almost uniform. This additional structure can sometimes be used to calculate bounds on $d(\mu^{k*}, U)$. Matthews [8] uses this technique to get a lower bound on variation distance for a random walk on the symmetric group generated by random transpositions. Here the strong uniform time will be used to give the result (1.5).

The coupling and strong uniform time used here are applicable to more general symmetric and some nonsymmetric random walks. The computations become more difficult but may still be useful. No random walks other than the simple one mentioned above will be discussed here.

2. The coupling. The coupling time T is given in an unusual way. The process Y, started in the uniform distribution, is used to define T. Then the sample paths of X_k, $k = 0, \cdots, T$, are constructed from those of Y_k, $k = 0, \cdots, T$. X will be shown to have the proper marginal distribution. What X does at time k, given $k \leqq T$, will depend on what Y does up to time T, making the joint process $\{X, Y\}$ non-Markovian.

Let Y_0 have the uniform distribution on Z_2^N. Create a mythical $(N + 1)$st coordinate that moves at the kth step if $Y_k = Y_{k-1}$. Denote the new process on Z_2^{N+1} by Y^*. If $|Y_0|$ is odd, let $Y_0^*(N + 1) = 1$. Otherwise let $Y_0^*(N + 1) = 0$. Thus $|Y_0^*|$ is even. Let A denote the set of coordinates i for which $Y_0^*(i) = 1$. Let T be the first time k for which exactly half of the A-coordinates of Y_k^* are zeros. T is the stopping time of interest.

Now X must be constructed. Let A_1 (A_0) be the set of A-coordinates for which Y_T^* is 1 (0), listed in order of increasing coordinate index. Note that $|A_1| = |A_0|$. Make a list of pairs of coordinates consisting of the first coordinates of A_0 and A_1, the second coordinates of A_0 and A_1, etc. Then at step k, if Y_k^* is obtained by moving coordinate i, then X_k^* is obtained by moving coordinate i if $k > T$ or, if $k \leqq T$,

$$\begin{array}{ll} i & \text{if } i \in A^c, \\ j & \text{if } i \in A \text{ and } i \text{ and } j \text{ are paired.} \end{array} \tag{2.1}$$

Call this transformation of coordinates ψ. Note that $\psi = \psi^{-1}$.

PROPOSITION 2.2. *T as given above is a coupling.*

Proof. First note that X^* and Y^*, and hence X and Y, match from time T on. X^* and Y^* always match in the A^c-coordinates of Y^*. At time T, Y_T^* is 1 on A_1 and 0 on A_0. Thus $Y_T^* - Y_0^*$ is 0 on A_1 and 1 on A_0. Since the A_1-coordinates of X^* move in the same way as do the A_0-coordinates of Y^*, X_T^* is 1 on A_1 and, similarly, 0 on A_0. Therefore $X_T^* = Y_T^*$. By (2.1) after time T, X^* and Y^* move the same coordinates, so $X_k^* = Y_k^*$ given $k \geqq T$.

Y^* has the proper marginal distribution by definition, so all that remains is to show that X^*, and hence X, has the proper marginal distribution.

Define a new process W^* by

$$\begin{aligned} &W_0^* = (0, \cdots, 0), \\ &\text{if } k \leqq T \quad m_k(W^*) = m_k(Y^*), \\ &\text{if } k > T \quad m_k(W^*) = \psi(m_k(Y^*)). \end{aligned}$$

W^* has the same distribution as $Y^* - Y_0^*$. It also agrees with $Y^* - Y_0^*$ up to time T. After time T, the moves of Y^* are permuted to obtain the moves of W^* by a transformation ψ that depends only on the past up to time T. Since moves are chosen independently and uniformly at each step, W^* is clearly a simple random walk on Z_2^{N+1}.

To show X^* is a simple random walk on Z_2^{N+1}, it suffices to show that for any $K > 0$ and $i_1, \cdots, i_K \in \{1, 2, \cdots, N+1\}$

$$P\left(\bigcup_{k=1}^{K} m_k(X^*) = i_k\right) = P\left(\bigcup_{k=1}^{K} m_k(W^*) = i_k\right) \tag{2.3}$$

X^* and W^* are related in a simple manner; if W_k^* moves coordinate i then X_k^* moves coordinate $\psi(i)$. Thus (2.3) is equivalent to

$$P\left(\bigcup_{k=1}^{K} m_k(W^*) = \psi(i_k)\right) = P\left(\bigcup_{k=1}^{K} m(W_k^*) = i_k\right). \tag{2.4}$$

Recall that A is the set of coordinates in which Y_0^* is one. Let S denote a subset of A of cardinality $|A|/2$. Note that T is the first time k such that W_k^* has $|A|/2$ ones in its A-coordinates.

Condition on $T = T^*, A = A^*$ and $\{A_0 = S\} \cup \{A_1 = S\}$. Given this, ψ is determined, and by the symmetry of W^* it is clear that $P(A_0 = S) = P(A_1 = S) = \frac{1}{2}$. There are exactly as many sample paths leading to $A_0 = S$ as to $A_1 = S$ and they can be put into one-to-one correspondence via ψ. Thus

$$\begin{aligned} &P\left(\bigcup_{k=1}^{K} m_k(W^*) = \psi(i_k) | T = T^*, A = A^*, \{A_0 = S\} \cup \{A_1 = S\}\right) \\ &\qquad = P\left(\bigcup_{k=1}^{K} m_k(W^*) = i_k | T = T^*, A = A^*, \{A_1 = S\} \cup \{A_0 = S\}\right). \end{aligned}$$

Summing over all possible values of T^*, A^* and S yields (2.4), hence (2.3), and X^* is a simple random walk on Z_2^{N+1}.

Now consider the distribution of T. Let Z_k denote the number of ones in the A-coordinates of Y_k^*. T is the first time k is such that $Z_k = Z_0/2$. The first time $|X_k^*| = (N+1)/2$ for $N+1$ even is an important hitting time for the Ehrenfest chain. Kemperman [6] gives the distribution of this hitting time; it is much like the result (1.5).

Given Z_0 and $T \leqq k$, the distribution of Z_k is symmetric about $Z_0/2$. Given Z_0 and $T > k$, the distribution of Z_k is concentrated to the right of $Z_0/2$. Thus

$$P(T \leqq k | Z_0) = P(Z_k = Z_0/2 | Z_0) + 2P(Z_k < Z_0/2 | Z_0). \tag{2.5}$$

Let $\lambda = \frac{1}{4}(\log N + c)$ and let κ be Poisson $((N+1)\lambda)$. At time κ each coordinate has moved a Poisson (λ) number of times, independently of how many times the other coordinates have moved. So each A-coordinate has probability

$$\sum_{i=1,3,5\cdots} \frac{e^{-\lambda}\lambda^i}{i!} = \frac{1}{2}\left(1 - \frac{e^{-c/2}}{\sqrt{N}}\right)$$

of being a 0 (moving an odd number of times) independently. Thus given Z_0, the number of zeros in the A-coordinates of Y^*_κ is binomial (Z_0, $\frac{1}{2}(e^{-c/2}/\sqrt{N})$). A normal approximation to the binomial and (2.5) give

$$P(T > \kappa | Z_0) = 2\Phi\left(e^{-c/2}\sqrt{\frac{Z_0}{N}}\right) - 1 + o(1). \tag{2.6}$$

The left side of (2.6) is

$$\sum_{k=0}^{\infty} P(T > k | Z_0) P(\kappa = k).$$

The facts that $P(T > k | Z_0)$ is a nonincreasing function of k, that the Poisson $((N+1)\lambda)$ puts asymptotically all its mass in the interval $(N+1)/4(\log N + c) \pm \sqrt{N} \log N$, and that the right side of (2.6) is unchanged when λ is changed by $O(\log N\sqrt{N})$ imply that

$$P(T > K | Z_0) = 2\Phi\left(e^{-c/2}\sqrt{\frac{Z_0}{N}}\right) - 1 + o(1).$$

Finally, consider the distribution of Z_0. Z_0 is approximately normal $(N/2, N/4)$, so as $N \to \infty$, Z_0 is between $N/2 - \sqrt{N} \log N$ and $N/2 + \sqrt{N} \log N$ with probability going to 1. Thus

$$P(T > K) = 2\Phi(e^{-c/2}/\sqrt{2}) - 1 + o(1),$$

verifying (1.9).

3. The strong uniform time. Consider the random walk X^* on Z_2^{N+1} as before and think of the $N+1$ coordinates as the vertices of a graph. At each step of the random walk an edge will be added to the graph. The first time, T, the graph is connected will be the strong uniform time of interest.

Edges will be assigned by the following mechanism. At time 0 there are no edges. Before step k, set up an $N+1 \times N+1$ matrix P^k with rows summing to 1 and nonnegative entries, such that $P^k_{ii} = 0$ for all i and $P^k_{ij} = P^k_{ji}$ for all i, j. The matrix chosen can depend on the past moves of the random walk only through the edges on the graph at time $k-1$. Then if the random walk moves coordinate i, pick another coordinate j with the probability distribution $(P^k_{i}, \cdots, P^k_{iN+1})$. Draw an edge connecting i and j. Since our interest is in whether the graph is connected, multiple edges may be ignored. Note that P (edge ij drawn at time $k | X^*_1, \cdots, X^*_{k-1}) = (P^k_{ij} + P^k_{ji})/(N+1) = 2P^k_{ij}/(N+1)$. Also,

$$\begin{aligned} &P(\text{coordinate } i \text{ moved} | \text{edge } ij \text{ drawn}) \\ &\quad = P(\text{coordinate } j \text{ moved} | \text{edge } ij \text{ drawn}) = \tfrac{1}{2}. \end{aligned} \tag{3.1}$$

This property will make T a strong uniform time.

PROPOSITION 3.2. *T is a strong uniform time for X.*

Proof. The proof uses induction on k. Let G_k be the graph at time k and let $E_1, \cdots, E_k$ be the sequence of edges added to make G_k. Finally let $z_i \in Z_2^{N+1}$ have a 1 in each coordinate that is a vertex of the edge E_i and zeros elsewhere. It will be shown by induction that

$$P(X_k^* = z|E_1 \cdots E_k) = P(X_k^* = z + z_i|E_1 \cdots E_k) \tag{3.3}$$

for all $i \leqq k$ and $z \in Z_2^{N+1}$. Repeated application of (3.3) shows that the distribution of X_k^* is invariant under addition of arbitrary combinations of $z_1 \cdots z_k$.

First note that (3.3) will imply the result. If G_k is connected, then $z_1 \cdots z_k$ generate the subgroup Z_E^{N+1} of Z_2^{N+1} consisting of all members of Z_2^{N+1} with an even number of ones. The conditional distribution of X_k^* given G_k connected is thus invariant under addition of members of Z_E^{N+1}. This implies that the marginal distribution of the first N coordinates of X_k^*, those making up X_k, is conditionally invariant under addition of all members of Z_2^N, and hence conditionally uniformly distributed on Z_2^N.

Result (3.3) is easily shown by induction. At time 0, there are no edges, so (3.3) is trivially true. Suppose (3.3) is true at time $k - 1$ and edge E_k is added at step k. Given G_{k-1} and E_k, $m_k(X^*)$ is equally likely to be either of the two coordinates connected by E_k from (3.1). Thus the conditional distribution of X_k^* given G_{k-1} and that E_k is invariant under addition of z_k. By induction the conditional distribution of X_{k-1}^* is invariant under addition of z_i for $i < k$. Since Z_2^{N+1} is abelian, the conditional distribution of X_k^* given $E_1 \cdots E_k$ is invariant under additions of $z_1 \cdots z_k$. So (3.3) is true by induction and Proposition 3.2 follows.

Intuitively one would like to choose the matrix P^k to give as much probability as possible to connecting disjoint components of G_{k-1}. However a simpler approach is sufficient to obtain (1.10). It is enough to let $P_{ij}^k = 1/N$ for all $i \neq j$ for all k. This corresponds to choosing two vertices at random at each step and connecting them. Each pair of vertices is chosen with probability $1/\binom{N+1}{2}$ at each step. After $J = (N/2)(\log N + b)$ steps, the expected number of distinct edges is at least

$$\frac{N}{2}(\log N + b) - \sum_{i=0}^{J} i \Big/ \binom{\mathrm{N}+1}{2} = \frac{N}{2}(\log N + b) - O(\log^2 N). \tag{3.4}$$

Thus, as $N \to \infty$, with probability approaching 1, there are at least

$$(N/2)(\log N + b - N^{-1/2})$$

distinct edges at time J. Theorem VII.3 of [3] implies that

$$P(T > J) = 1 - e^{-e^{-b}} + o(1),$$

which gives the upper bound (1.10).

As discussed in the Introduction, separation is often mainly of interest as an upper bound on variation distance. Using the structure of a random graph with $K = (N/4)(\log N + c + o(1))$ edges, we can deduce the result (1.4) and thus calculate the variation distance as in (1.5). The following is a sketch of the necessary calculations.

Consider the components of the graph at time K. Theorems 2 and 4 (Chapters 1 and 7, respectively) of Kolchin, Sevast'yanov and Chistyakov [7] imply that the number of isolated vertices at time K is approximately binomial $(N, e^{-c/2}/\sqrt{N})$. Also, as in Theorem VII.2 of Bollobas [3], there are $O_p(\log N)$ coordinates in small components of the graph, and the remainder are in the giant component. These $O_p(\log N)$ coordinates can be ignored; their being all zeros or all ones will not affect the results. Also ignore one coordinate of the giant component. Then, as in the proof of Proposition 3.2, the remaining coordinates of the giant component are 0 or 1 with probability $\frac{1}{2}$ each, independently.

For these coordinates and the isolated vertices, to generate essentially the same distribution of number of ones, it would suffice to let each coordinate be one with probability $\frac{1}{2}(1 - e^{-c/2}/\sqrt{N})$.

From this a result similar to (1.4) sufficient to derive (1.5) follows.

REFERENCES

[1] D. ALDOUS AND P. DIACONIS, *Shuffling cards and stopping times*, Amer. Math. Monthly, 93 (1986), pp. 333–348.

[2] ———, *Strong uniform times and finite random walks*, Technical Report No. 59, Dept. of Statistics, University of California, Berkeley, CA, 1986.

[3] B. BOLLOBAS, *Random Graphs*, Academic Press, New York, 1985.

[4] P. DIACONIS, *Group Theory in Statistics*, IMS, Hayward, CA, 1987, to appear.

[5] D. GRIFFEATH, *A maximal coupling for Markov chains*, Z. Wahrsch. Verw. Gebiete, 31 (1975), pp. 95–106.

[6] J. KEMPERMAN, *The Passage Problem for a Stationary Markov Chain*, University of Chicago Press, Chicago, IL, 1961.

[7] V. KOLCHIN, B. SEVAST'YANOV AND V. CHISTYAKOV, *Random Allocations*, V. H. Winston, Washington, DC, 1978.

[8] P. MATTHEWS, *A strong uniform time for random transpositions*, Technical Report No. 86, Dept. of Statistics, Purdue University, West Lafayette, IN, 1986.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

019

AN ALGEBRAIC CONSTRUCTION OF SONAR SEQUENCES USING *M*-SEQUENCES*

RICHARD A. GAMES†

Abstract. An algebraic construction of sonar sequences that is based on the properties of q-ary M-sequences for q a prime power is presented. Sonar sequences give two-dimensional synchronization patterns that have two-dimensional spatial aperiodic autocorrelation functions with minimum out-of-phase values. The best sonar sequences with length $q^m \leqq 128$ that are obtained from the construction are tabulated. Based on a comparison with the limited number of known optimal values, the construction performs quite well, producing optimal sonar sequences in the majority of applicable cases.

Key words. M-sequences, sonar sequences, two-dimensional synchronization

AMS(MOS) subject classifications. 05B, 94A

1. Introduction. This paper describes an algebraic construction of sonar sequences that is based on the properties of q-ary M-sequences—maximum period linear recursive sequences over GF(q) for a prime power q. An M by N *sonar sequence* $a_1, a_2, \cdots, a_N$ has integer values in the range 0 to $M-1$ and satisfies the *synchronization* property: for each k from 1 to $N-1$, the list of differences $(a_{i+k} - a_i\colon i = 1, 2, \cdots, N-k)$ contains distinct entries.

A sonar sequence can be pictured as an M by N *sonar array* (with rows and columns numbered 0 to $M-1$ and 1 to N, respectively) in which column i has a single dot in row a_i. The synchronization property is equivalent to the fact that any horizontal and/or vertical shifted copy of the M by N array will agree with the original in at most one dot [5]. In other words, the two-dimensional spatial *aperiodic autocorrelation* function of the array has out-of-phase values of at most 1. Figure 1 shows a 4 by 8 sonar sequence and array.

In applications, the sonar sequence corresponds to a sequence of transmitted tones; a horizontal shift of the array corresponds to elapsed time; a vertical shift corresponds to a Doppler shift in frequency. The synchronization property guarantees that out-of-phase shifts result in "correlations" of at most one, which is small compared to the matched value of N. See [2], [5], [6] for more on these and other related two-dimensional synchronization patterns.

The fundamental problem for sonar sequences is to determine for fixed M the largest value of N for which there exists an M by N sonar sequence. It is not hard to see that with M rows, the maximum number of columns is at most $2M$. However the following known optimal values, given in [8], indicate that as M increases the maximum value of N is probably closer to M than $2M$: 1×2, 2×4, 3×6, 4×8, 5×9, 6×11, 7×12, 8×13, 9×14, 10×16, 11×17 and 12×18.

This paper contains an algebraic construction that produces for q a prime power and m a positive integer, a sonar sequence with q^m columns. Initially, the sonar sequence produced has $q^m - 1$ rows; however, the sequence satisfies a stronger synchronization property that allows the rows of the sonar array to be rotated cyclically while still preserving the synchronization property. Thus, a better sonar sequence for this case can be obtained by rotating empty rows to the top (or bottom) and deleting them. The decrease in the

* Received by the editors February 18, 1986; accepted for publication (in revised form) June 8, 1987. This work was supported by the MITRE Sponsored Research Program. A version of this paper was presented at the IMA Conference on Cryptography and Coding, Cirencester, England, December 15–17, 1986.

† The MITRE Corporation, Bedford, Massachusetts 01730.

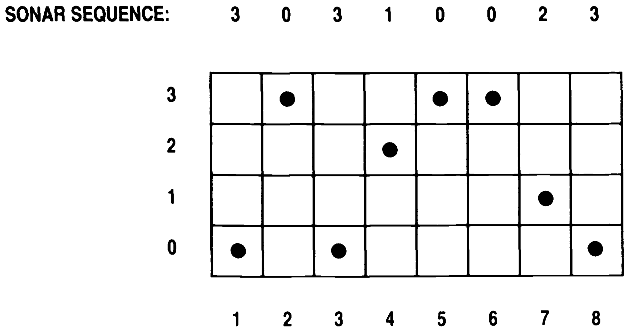


FIG. 1. *A* 4 *by* 8 *sonar array.*

number of rows from $q^m - 1$ that can be obtained depends on the primitive polynomial used in the construction. A further decrease, which depends on two strong-synchronization preserving transformations, is sometimes possible. The parameters of the best sonar sequences obtained from the construction are listed in Table 2 for all cases q and m with $q^m \leqq 128$. Based on a comparison with the limited number of known optimal values, the construction performs quite well, producing optimal sonar sequences in the majority of applicable cases.

2. *M*-sequences and shift sequences. The construction for sonar sequences is based on the properties of q-ary M-sequences. A q-ary *M-sequence s of span n* and period $q^n - 1$ is determined by choosing a primitive polynomial $f(x)$ over GF(q) of degree n. The sequence is formed by choosing initial conditions $s_0, s_1, \cdots, s_{n-1}$ and generating the sequence using the linear recursion that has characteristic polynomial $f(x)$. The sequence is denoted by $s = (s_0, s_1, \cdots, s_{q^n-2})$, and s is identified with its $q^n - 1$ cyclic shifts $E^k s$, $k = 0, 1, \cdots, q^n - 2$, where E is the sequence *shift operator*; i.e., Es is the sequence with ith term $(Es)_i = s_{i+1}$. Each cyclic shift of s corresponds to a distinct choice of initial conditions.

It is well known [1], [7] that if m divides n (so that $q^m - 1$ divides $q^n - 1$) and if the sequence $s = (s_0, s_1, \cdots, s_{q^n-2})$ is arranged in a $(q^m - 1)$ by $v = (q^n - 1)/(q^m - 1)$ array:

$$A(s) = \begin{bmatrix} s_0 & s_1 & \cdots & s_{v-1} \\ s_v & s_{v+1} & \cdots & s_{2v-1} \\ & & \vdots & \\ s_{(q^m-2)v} & s_{(q^m-2)v+1} & \cdots & s_{(q^m-1)v-1} \end{bmatrix},$$

then each column of this array is either identically zero or a shift of the same M-sequence of span m. If the column sequence is denoted by t, then the ith column of $A(s)$ is either identically 0 or has the form $E^{e_i}t$ for some integer e_i with $0 \leqq e_i \leqq q^m - 2$. Using the convention that $E^{\infty}t = 0$, i.e., $e_i = \infty$ if column i is identically zero, we obtain the corresponding sequence $(e_0, e_1, \cdots, e_{v-1})$ for the M-sequence s.

A sequence e of period $q^n - 1$ is defined by the array

$$A(e) = \begin{bmatrix} e_0 & e_1 & \cdots & e_{v-1} \\ e_0 - 1 & e_1 - 1 & \cdots & e_{v-1} - 1 \\ & & \vdots & \\ e_0 - (q^m - 2) & e_1 - (q^m - 2) & \cdots & e_{v-1} - (q^m - 2) \end{bmatrix}.$$

Here we use the convention $\infty - i = \infty$. For example, the entries in row 2 of $A(e)$

correspond to the shifts of the column sequence involved in the array $A(E^v s)$. The finite elements of $A(e)$ are regarded as elements of $Z/(q^m - 1)$, the integers modulo $q^m - 1$, to obtain entries in the range $0 \leqq e_i \leqq q^m - 2$. For a fixed integer m dividing n, the sequence e of period $q^n - 1$ is determined, up to cyclic shifts, by the primitive polynomial $f(x)$ and is called the *shift sequence associated with* $f(x)$ *and* m. The difference properties of this shift sequence are used in the sonar sequence construction.

3. The sonar sequence construction. Let q be a prime power and let $m \geqq 1$ be an integer. The construction uses a q-ary M-sequence of span $2m$. In this case $v = (q^{2m} - 1)/(q^m - 1) = q^m + 1$. Let $f(x)$ be a primitive polynomial of degree $2m$ over GF(q), and let $e = (e_0, e_1, \cdots, e_{q^{2m}-2})$ be the associated shift sequence. For $a \in Z/(q^m - 1)$, we use the convention that $a - \infty = \infty - a = \infty$.

The following facts are special cases of results proved in [4]. (The results in [4] are stated for the case $q = 2$; however, the proofs remain valid for q any prime power.)

FACT 1. In any v consecutive terms of e, there is exactly one ∞ [4, Thm. 1].

FACT 2. For fixed $k \in Z/(q^m - 1)$, $k \not\equiv 0 \pmod{v}$, the list of differences $(e_{i+k} - e_i$: $i = 0, 1, \cdots, v - 1)$ contains each element of $Z/(q^m - 1)$ exactly once [4, Thm. 2].

If $e = (e_0, e_1, \cdots, e_{q^{2m}-2})$ is the shift sequence associated with a primitive polynomial over GF(q) of degree $2m$, then e can be shifted so that $e_0 = \infty$. Then Fact 1 implies that $e_1, e_2, \cdots, e_{q^m}$ are all elements of $Z/(q^m - 1)$. Furthermore, Fact 2 implies that the sequence $e_1, e_2, \cdots, e_{q^m}$ has the synchronization property, since certainly, for $1 \leqq k \leqq q^m - 1$, the differences $(e_{i+k} - e_i$: $i = 1, 2, \cdots, q^m - k)$ being distinct modulo $q^m - 1$ means they are distinct as integers. Thus $e_1, e_2, \cdots, e_{q^m}$ forms a $(q^m - 1)$ by q^m sonar sequence.

However, more is true. Consider the sequence $f_1 = e_1 - 1$, $f_2 = e_2 - 1, \cdots, f_{q^m} = e_{q^m} - 1$, where each entry is considered modulo $q^m - 1$. The differences $(f_{i+k} - f_i$: $i = 1, 2, \cdots, q^m - k)$ do not change modulo $q^m - 1$, and so $f_1, f_2, \cdots, f_{q^m}$ is also a $(q^m - 1)$ by q^m sonar sequence, which corresponds to the sequence $E^v s$. The new sonar array is formed by cyclically rotating the rows of the former sonar array down by one. This, of course, can be repeated.

In general, the sequence $a_1, a_2, \cdots, a_N$ of integers in the range 0 to $M - 1$ satisfies the *strong synchronization* property if for each k from 1 to $N - 1$, the list of differences $(a_{i+k} - a_i \pmod{M}$: $i = 1, 2, \cdots, N - k)$ contains distinct entries. As has already been seen, a sonar sequence with the strong synchronization property can be "rotated" to yield another sonar sequence. In general, this is not the case. If an M by N sonar array with the strong synchronization property has d consecutive empty rows, then an improved $M - d$ by N sonar array can be obtained by rotating these empty rows to the bottom and deleting them. The resulting array can no longer have the strong synchronization property.

Example 1. $q = 2$, $m = 3$.

The polynomial $f(x) = x^6 + x^5 + x^2 + x + 1$ is primitive over GF(2). Starting with the state (0, 0, 0, 0, 0, 1) the following binary M-sequence s is obtained (arranged as a 7 by $v = (2^6 - 1)/(2^3 - 1) = 9$ array):

$$A(s) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The column M-sequence is 0010111 (taken from the first column). The shifts of this sequence occurring in the other columns determine the first 9 terms of the shift sequence:

$$0, \infty, 6, 0, 0, 2, 5, 2, 1.$$

The remaining terms of the shift sequence continue as

$$|6, \infty, 5, 6, 6, 1, 4, 1, 0|5, \infty, 4, 5, 5, 0, 3, 0, 6|4, \infty, 3, \cdots.$$

Shift the ∞ term to the beginning and the next 8 terms form the 7×8 sonar sequence:

$$6, 0, 0, 2, 5, 2, 1, 6.$$

The differences for this sequence are:

$$\begin{array}{ccccccccccccc}
-6 & & 0 & & 2 & & 3 & & -3 & & -1 & & 5 \\
 & -6 & & 2 & & 5 & & 0 & & -4 & & 4 & \\
 & & -4 & & 5 & & 2 & & -1 & & 1 & & \\
 & & & -1 & & 2 & & 1 & & 4 & & & \\
 & & & & -4 & & 1 & & 6 & & & & \\
 & & & & & -5 & & 6 & & & & & \\
 & & & & & & 0 & & & & & &
\end{array}$$

The array corresponding to this sonar sequence is shown in Fig. 2. Figure 3 shows the sonar array obtained by rotating the empty rows to the bottom and deleting. The corresponding sonar sequence is 1, 2, 2, 4, 0, 4, 3, 1. To obtain the best sonar sequence for this case, all primitive polynomials of degree 6 over GF(2) must be considered. The next section describes how this can be done by just considering decimations.

4. Sonar sequences obtained from decimations. All q-ary M-sequences of span n can be obtained from a single q-ary M-sequence of span n by decimating by integers r with $(r, q^n - 1) = 1$. For a fixed integer m dividing n, the associated shift sequences can be obtained by decimations along with one additional multiplication. For a sequence $s = (s_0, s_1, \cdots, s_{q^n-2})$ and an integer r, the *r-decimation* of s is denoted by $s[r]$ and has ith term $(s[r])_i = s_{ri}$.

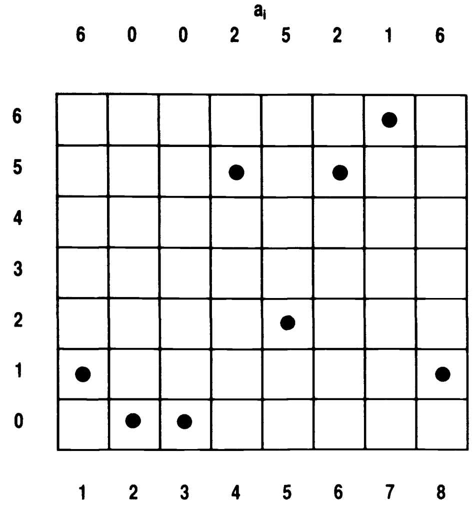


FIG. 2. *A* 7 *by* 8 *sonar array.*

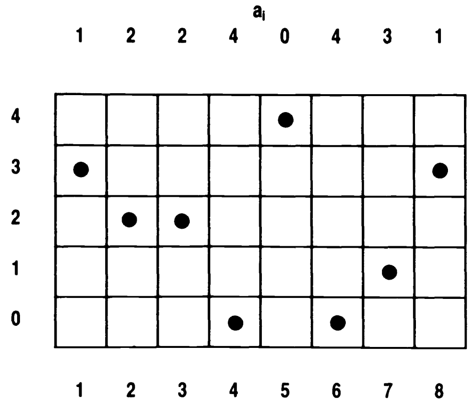


FIG. 3. *A* 5 *by* 8 *sonar array.*

THEOREM 1. *Let m be an integer dividing n and let* $e = (e_0, e_1, \cdots, e_{q^n-2})$ *be the shift sequence associated with m and a primitive polynomial* $f(x)$ *over* GF(q) *of degree n, equivalently, with the M-sequence* $s = (s_0, s_1, \cdots, s_{q^n-2})$ *generated by* $f(x)$. *If r is an integer with* $(r, q^n - 1) = 1$, *then the shift sequence* $f = (f_0, f_1, \cdots, f_{q^n-2})$ *associated with m and the M-sequence* $s[r]$ *satisfies*

$$f_i = r^{-1}e_{ri}(\text{mod } q^m - 1).$$

Proof. Since $q^m - 1$ divides $q^n - 1$, $(r, q^n - 1) = 1$ implies $(r, q^m - 1) = 1$, and so $r^{-1}(\text{mod } q^m - 1)$ exists. If s is shifted so that $s_0 \neq 0$, then the first column of $A(s)$ can be taken as the column sequence for $A(s)$. This sequence is $t = (t_0, t_1, \cdots, t_{q^m-2})$ with jth term $t_j = s_{jv}$. The first column of $A(s[r])$ is $(s_{0r}, s_{vr}, s_{2vr}, \cdots, s_{(q^m-2)vr})$, which is $t[r]$.

By definition of the shift sequence $E^{e_i}(s_i, s_{i+v}, s_{i+2v}, \cdots, s_{i+(q^m-2)v}) = t$, i.e., for $i = 0, 1, \cdots, q^n - 2, j = 0, 1, \cdots, q^m - 2$,

$$s_{jv} = s_{i+(j+e_i)v} = s_{i+jv+e_iv}. \tag{1}$$

The proof involves writing the indices in (1) in terms of the decimation value r. For fixed i, j and e_i, since $(r, q^n - 1) = (r, q^m - 1) = 1$, i', j' and $f_{i'}$ can be determined so that $i = ri'(\text{mod } q^n - 1)$, $j = rj'(\text{mod } q^m - 1)$ and $e_i = rf_{i'}(\text{mod } q^m - 1)$. Note that $\{i': i = 0, 1, \cdots, q^n - 2\} = \{0, 1, \cdots, q^n - 2\}$ and $\{j': j = 0, 1, \cdots, q^m - 2\} = \{0, 1, \cdots, q^m - 2\}$. So for $i' = 0, 1, \cdots, q^n - 2, j' = 0, 1, \cdots, q^m - 2$, substituting into (1),

$$s_{vrj'} = s_{ri' + vrj' + vrf_{i'}} = s_{r(i' + v(j' + f_{i'}))}$$

or equivalently

$$t[r] = E^{f_{i'}}(s_{ri'}, s_{r(i'+v)}, s_{r(i'+2v)}, \cdots, s_{r(i'+(q^m-2)v)})$$
$$= E^{f_{i'}}(s[r]_{i'}, s[r]_{i'+v}, \cdots, s[r]_{i'+(q^m-2)v}).$$

Thus $f = (f_0, f_1, \cdots, f_{q^n-2})$ is the shift sequence for $s[r]$ where for $i = 0, 1, \cdots, q^n - 2$, $f_i = r^{-1}e_{ri}(\text{mod } q^m - 1)$. □

Theorem 1 can be used to obtain sonar sequences for each $r \in Z^*_{q^{2m}-1} = \{i: 0 \leqq i \leqq q^{2m} - 2, (i, q^{2m} - 1) = 1\}$. Actually, it suffices to decimate by a set of representatives of the cosets of $Z^*_{q^{2m}-1}/\{1, q, q^2, \cdots, q^{2m-1}\}$, since elements of the same coset correspond to the same primitive polynomial. Example 2 illustrates the process for the sequence considered in Example 1.

Example 2. $q = 2, m = 3$.

From Example 1, the shift sequence of the primitive polynomial $f(x) = x^6 + x^5 + x^2 + x + 1$ is

$$A(e) = \begin{bmatrix} \infty & 6 & 0 & 0 & 2 & 5 & 2 & 1 & 6 \\ \infty & 5 & 6 & 6 & 1 & 4 & 1 & 0 & 5 \\ \infty & 4 & 5 & 5 & 0 & 3 & 0 & 6 & 4 \\ \infty & 3 & 4 & 4 & 6 & 2 & 6 & 5 & 3 \\ \infty & 2 & 3 & 3 & 5 & 1 & 5 & 4 & 2 \\ \infty & 1 & 2 & 2 & 4 & 0 & 4 & 3 & 1 \\ \infty & 0 & 1 & 1 & 3 & 6 & 3 & 2 & 0 \end{bmatrix}.$$

To obtain other examples, decimate by $r \in Z_{63}^*$. The cosets of $Z_{63}^*/\{1, 2, 4, 8, 16, 32\}$ are as shown in Table 1. The sequence $e[11]$ begins

$$\infty, 1, 0, 1, 5, 0, 0, 3, 1, \infty, \cdots$$

so that the shift sequence $(11^{-1})e[11] = 2e[11]$ begins

$$\infty, 2, 0, 2, 3, 0, 0, 6, 2, \infty, \cdots.$$

The sonar sequence is 2, 0, 2, 3, 0, 0, 6, 2, which also results in a 5×8 sonar array when the two consecutive empty rows are rotated to the bottom and deleted. Decimating by 5, 13, 23 and 31 similarly produce 5×8 arrays, although in general some decimation values will result in a different number of rows.

5. Two strong-synchronization preserving transformations. There are two transformations that can be applied to a sequence which preserve the strong synchronization property: multiplication and shearing. The observation that shearing could be useful in this context is due to O. Moreno.

If $a_1, a_2, \cdots, a_N$ is an M by N sonar sequence with the strong synchronization property and r is an integer with $(r, M) = 1$, then $ra_1, ra_2, \cdots, ra_n(\text{mod } M)$ also has the strong synchronization property. Multiplication by r simply permutes the rows of the sonar array and the differences in the difference triangle. The multiplied sequence can correspond to a sonar array with more consecutive empty rows, and a better sonar array can be obtained for this case.

Example 3. $q = 2, m = 3$.

The sonar sequence 2, 0, 2, 3, 0, 0, 6, 2 of Example 2 when multiplied by $r = 2$ becomes 4, 0, 4, 6, 0, 0, 5, 4. This latter sequence corresponds to a sonar array with empty rows at 1, 2 and 3. Thus a 4 by 8 sonar sequence, which is optimal, can be obtained in this case. This sonar array is pictured in Fig. 1 and can be computed directly using the shift sequence of the primitive polynomial $f(x) = x^6 + x^5 + 1$ with multiplier $r = 2$.

TABLE 1

Cosets	Representative r	r^{-1}(mod 7)
{1, 2, 4, 8, 16, 32}	1	1
{5, 10, 20, 40, 17, 34}	5	3
{11, 22, 44, 25, 50, 37}	11	2
{13, 26, 52, 41, 19, 38}	13	6
{23, 46, 29, 58, 53, 43}	23	4
{31, 62, 61, 59, 55, 47}	31	5

Unlike the multiplication transformation, which simply permutes empty rows, the shearing transformation [6] can increase the number of empty rows. If $a_1, a_2, \cdots, a_N$ is an M by N sonar sequence with the strong synchronization property, and s is an integer with $0 \leqq s \leqq M - 1$, then the *sheared* sequence $a_1, a_2 + s, \cdots, a_N + (N - 1)s(\text{mod } M)$ also has the strong synchronization property. Shearing by s adds the constant s to each term of the difference triangle, permuting these entries (mod M).

Example 4. $q = 2, m = 3$.

The sonar sequence 1, 0, 0, 5, 2, 5, 6, 1 of Example 1 when sheared by $s = 2$ becomes 1, 2, 4, 4, 3, 1, 4, 1. This latter sequence corresponds to a sonar array with empty rows at 0, 5 and 6. Thus a 4 by 8 sonar sequence, which is optimal, can be obtained in this case. It is the mirror image of the sonar array pictured in Fig. 1.

6. The best sonar sequences obtained from the construction. When tabulating the best sonar sequences obtained from the construction, the next theorem implies that only polynomials over GF(p), p a prime, need to be considered.

THEOREM 2. *Let m and r be positive integers and $q = p^r$,, p a prime. Let $e = (e_0, e_1, \cdots, e_{q^{2m}-2})$ be the shift sequence associated with m and a primitive polynomial $f(x)$ over* GF(q) *of degree $2m$. Then there exists a primitive polynomial $g(x)$ over* GF(p) *of degree $2mr$ such that e is the shift sequence associated with mr and $g(x)$.*

Proof. In this case $v = q^m + 1 = p^{mr} + 1$. The polynomial $f(x)$ can be used to construct $\text{GF}(q^{2m}) \cong \text{GF}(p^{2mr})$ where a root α of $f(x)$ can be taken as a primitive element. Let $g(x)$ be the minimal polynomial of α over GF(p); i.e., $g(x) = (x - \alpha)(x - \alpha^p) \cdots (x - \alpha^{p^{2mr-1}})$. Then $g(x)$ is a primitive polynomial over GF(p) of degree $2mr$. To see that the shift sequence associated with $f(x)$ and m is identical to the shift sequence associated with $g(x)$ and mr, the trace function definition of M-sequences is used. The *trace function* $\text{Tr}_q^n : \text{GF}(q^n) \rightarrow \text{GF}(q)$ is defined for $x \in \text{GF}(q^n)$ by $\text{Tr}_q^n(x) = x + x^q + \cdots + x^{q^{n-1}}$. The M-sequence $s = (s_0, s_1, \cdots, s_{q^{2m}-2})$ generated by $f(x)$ is determined, up to cyclic shift, by $s_i = \text{Tr}_q^{2m}(\alpha^i)$, $i = 0, 1, \cdots, q^m - 2$. Similarly, the M-sequence $u = (u_0, u_1, \cdots, u_{p^{2mr}-2})$ generated by $g(x)$ is determined, up to cyclic shift, by $u_i = \text{Tr}_p^{2mr}(\alpha^i)$, $i = 0, 1, \cdots, p^{2mr} - 2$. However,

$$\text{Tr}_p^{2mr}(x) = \text{Tr}_p^r(\text{Tr}_{p^r}^{2m}(x)),$$

and so $u_i = \text{Tr}_p^r(s_i)$, $i = 0, 1, \cdots, p^{2mr} - 2$. In other words, the $p^{mr} - 1$ by v array $A(u)$ is obtained by applying the function $\text{Tr}_p^r : \text{GF}(q) \rightarrow \text{GF}(p)$ to each term of the $q^m - 1$ by v array $A(s)$. Thus, the column shifts involved in each array are identical; i.e., the shift sequence associated with $f(x)$ and m is identical to the shift sequence associated with $g(x)$ and rm. □

Table 2 contains the parameters of the best sonar sequences with 128 or fewer columns that can be obtained from the construction. For each prime p and integer m with $p^m \leqq 128$, a primitive polynomial over GF(p) of degree $2m$, a multiplier r and a shear factor s used to obtain a sonar sequence with these parameters are listed. Table 2 lists these parameters with the number of columns in increasing order.

The optimal sonar sequence parameters are known for 18 or fewer columns. For a fixed number of columns N, the minimum number of rows M possible can be determined from the results of [8] given in § 1. There are 11 prime powers less than 18 for which the present construction applies. Of these 11 cases, eight have optimal parameters (1 × 2, 2 × 3, 2 × 4, 3 × 5, 4 × 7, 4 × 8, 10 × 16, and 11 × 17), while the remaining three cases have only one extra row (6 × 9, 7 × 11 and 9 × 13).

7. Conclusion. An algebraic construction of sonar sequences with q^m columns, q a prime power and m an integer, was presented. The construction was based on the prop-

TABLE 2

Parameters of the best $M \times N$ sonar sequences obtained from the construction; $N \leqq 128$.

q	m	M	$N = q^m$	$N - M$	Primitive polynomial	r	s
2	1	1	2	1	1 1 1	1	0
3	1	2	3	1	1 1 2	1	0
2	2	2	4	2	1 1 0 1	1	0
5	1	3	5	2	1 1 2	1	0
7	1	4	7	3	1 1 3	1	0
2	3	4	8	4	1 1 0 0 0 0 1	2	0
3	2	6	9	3	1 1 0 0 2	1	0
11	1	7	11	4	1 1 7	1	0
13	1	9	13	4	1 11 6	1	0
2	4	10	16	6	1 0 1 1 0 0 1 0 1	4	0
17	1	11	17	6	1 8 6	3	0
19	1	13	19	6	1 11 3	1	0
23	1	16	23	7	1 22 19	9	0
5	2	19	25	6	1 1 0 2 3	5	0
3	3	21	27	6	1 0 2 0 1 1 2	9	0
29	1	22	29	7	1 7 2	9	0
31	1	23	31	8	1 24 17	11	0
2	5	24	32	8	1 0 0 0 0 1 0 0 1 1 1	2	0
37	1	28	37	9	1 31 15	13	13
41	1	32	41	9	1 32 7	19	0
43	1	34	43	9	1 10 5	19	0
47	1	38	47	9	1 19 22	21	0
7	2	40	49	9	1 3 3 2 3	23	0
53	1	44	53	9	1 30 33	3	0
59	1	49	59	10	1 49 11	15	0
61	1	51	61	10	1 3 54	7	25
2	6	54	64	10	1 1 1 0 0 0 0 0 0 0 1 0 1	17	0
67	1	57	67	10	1 4 2	5	0
71	1	60	71	11	1 14 42	29	32
73	1	62	73	11	1 48 31	19	0
79	1	67	79	12	1 54 63	31	0
3	4	71	81	10	1 1 2 0 2 2 1 0 2	9	0
83	1	72	83	11	1 3 24	27	0
89	1	77	89	12	1 15 58	17	0
97	1	85	97	12	1 52 39	17	0
101	1	88	101	13	1 28 63	27	0
103	1	91	103	12	1 39 86	29	0
107	1	94	107	13	1 33 97	25	0
109	1	95	109	14	1 2 40	29	34
113	1	100	113	13	1 30 54	33	0
11	2	109	121	12	1 3 2 10 2	43	0
5	3	112	125	13	1 3 4 3 0 3 2	59	0
127	1	114	127	13	1 2 23	41	0
2	7	117	128	13	1 1 1 1 0 1 0 1 0 1 1 0 1 1 1	45	0

erties of the shift sequence obtained from a q-ary M-sequence of span $2m$. The best sonar sequences obtained from the construction were tabulated for the number of columns $q^m \leqq 128$. Based on a comparison with the limited number of known optimal values, the construction performed quite well, producing optimal sonar sequences in eight out of the 11 applicable cases and being off by 1 in the remaining three cases.

A subject for future research is the asymptotic performance of the construction, including a comparison with the parabolic construction of [3]. Also, the results suggest that it should be possible to improve the upper bound of $2M$ on the number N of columns of an $M \times N$ sonar sequence. H. Taylor [9] has reported progress on this problem.

Acknowledgments. The author is indebted to Curtis P. Brown, Michael S. Chao, Robert A. Meyer and Robert D. Silverman for providing the programming support required to generate Table 2.

REFERENCES

[1] L. D. BAUMERT, *Cyclic Difference Sets*, Lecture Notes in Mathematics 182, Springer, Berlin, 1971.

[2] J. P. COSTAS, *A study of a class of detection waveforms having nearly ideal range-Doppler ambiguity properties*, IEEE Proceedings, vol. 72, August 1984, pp. 996–1009.

[3] R. GAGLIARDI, J. ROBBINS AND H. TAYLOR, *Acquisition sequences in PPM communications*, Dept. Elec. Engrg., Univ. of Southern California, Los Angeles, CA; IEEE Trans. Inform. Theory, submitted.

[4] R. A. GAMES, *Crosscorrelation of M-sequences and GMW-sequences with the same primitive polynomial*, Discrete Appl. Math., 12 (1984), pp. 139–146.

[5] S. W. GOLOMB AND H. TAYLOR, *Two-dimensional synchronization patterns for minimum ambiguity*, IEEE Trans. Inform. Theory, IT-28, (1982), pp. 600–604.

[6] ———, *Constructions and properties of Costas arrays*, IEEE Proceedings, vol. 72, September 1984, pp. 1143–1163.

[7] B. GORDON, W. H. MILLS AND L. R. WELCH, *Some new difference sets*, Canad. J. Math., 14 (1962), pp. 614–625.

[8] J. ROBBINS AND H. TAYLOR, *Sonar sequences and PPM sequences, part* 1, CSI-84-12-01, Communication Sciences Institute, Univ. Southern California, Los Angeles, CA, December, 1984.

[9] H. TAYLOR, personal communication.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

020

SOME COMPLETENESS RESULTS ON DECISION TREES AND GROUP TESTING*

DING-ZHU DU† AND KER-I KO‡

Abstract. The computational complexity of the group testing problem is investigated under the minimax measure and the decision tree model. We consider the generalizations of the group testing problem in which partial information about the decision tree of the problem is given. Using this approach, we demonstrate the NP-hardness of several decision problems related to various models of the group testing problem. For example, we show that, for several models of group testing, the problem of recognizing a set of queries that uniquely determines each object is co-NP-complete.

Key words. group testing, decision trees, NP-completeness, #P-completeness

AMS(MOS) subject classifications. 68Q25, 68P10

1. Introduction. Many combinatorial search problems involve the minimization of the heights of decision trees. Such problems can often be described as two-person query games, where one player A selects an object x from a finite domain D and assumes the role of an oracle while the other player B tries to identify the object x by making queries to A about the object. Consider, as an example, the problem of group testing [3], [7], [14], [23]–[26], [29]. The domain of the problem is the set $\mathcal{S}_{n,d}$ of all subsets of $\{1, \cdots, n\}$ that have size d. The player B tries to identify a set $S \in \mathcal{S}_{n,d}$ by making queries about S. Each query is a subset $T \subseteq \{1, \cdots, n\}$ and its answer, provided by A, is either "YES" if the intersection $S \cap T$ is nonempty, or "NO" otherwise. As another example, we may consider the problem of sorting by decision tree as a two-person query game [16], in which a domain consists of all permutations over $\{1, \cdots, n\}$ and, to identify a permutation α, queries of the form "$\alpha(i) < \alpha(j)$?" may be asked. An algorithm for such a search problem is essentially a general procedure to produce, for each domain, a decision tree of which each path uniquely determines an object in the domain. An optimal algorithm is one which produces, for each domain, a decision tree of the minimum height. For example, for the problem of group testing, a decision tree may be described as follows: Each node of the tree is a subset $T \subseteq \{1, \cdots, n\}$, and has two children, identified by answers YES and NO to the query T. Each path of the tree consists of a sequence of queries $(T_1, \cdots, T_m)$ with their answers $(a_1, \cdots, a_m)$ such that there is exactly one $S \in \mathcal{S}_{n,d}$ having the property that $S \cap T_i$ is nonempty if and only if $a_i =$ YES for $i = 1, \cdots, m$.

Except for a very few simple search problems, the problem of finding an optimal algorithm for a shortest decision tree problem appears to be intractable. For example, in spite of extensive studies, the optimal algorithms for sorting and group testing problems remain as open questions (cf. [20]). For the group testing problems, people have conjectured that they are indeed intractable [12]; however, no formal proofs for these conjectures have been found.

* Received by the editors May 27, 1986; accepted for publication July 23, 1987. A preliminary version of this paper was presented at the Third SIAM Conference on Discrete Mathematics, Clemson University, Clemson, South Carolina, May 1986.

† Mathematical Sciences Research Institute, Berkeley, California 94720. Current address, Institute of Applied Mathematics, Academia Sinica, Beijing, China.

‡ Department of Mathematics, University of California at Santa Barbara, Santa Barbara, California 93106. Current address, Department of Computer Science, SUNY at Stony Brook, Stony Brook, New York 11794. The research of this author was supported in part by the National Science Foundation under grants DCR83-12472 and DCR-8501226.

In the study of computational complexity of combinatorial optimization problems, a search problem is usually formulated as a decision problem so that the lower bound results are easier to be developed (often through the reductions from known NP- or PSPACE-complete problems). For the shortest decision tree problem, the associated decision problem may be formulated as follows:

> Given a domain D and an integer k, determine whether there is a decision tree of height $\leqq k$ of which each path uniquely determines an object in D.

It is not hard to see that the above problem is often solvable in polynomial space. (For given D and k, we may guess nondeterministically a decision tree of height k and verify that for each of its path, there is only one object consistent with the queries and answers of this path. Note that at any step of the computation, this algorithm needs only $O(k)$ space to store one path of the decision tree, although the complete tree contains about 2^k many nodes.) On the other hand, the domain of the problem often has a very simple form so that it is difficult to obtain a reduction from other (PSPACE-)complete problems to it since such a reduction would usually require rich structures in the problem in question (cf. [5], [8]). Indeed, it follows from the research in abstract complexity theory that if the input to a problem may be defined by two integers (here, n and d), then the problem cannot be PSPACE-complete unless P = PSPACE [6]. So, in order to obtain any completeness results on the shortest decision tree problems, we must reformulate the problems to add more complex structures to the problem instances. A general approach to this is to treat the problem as a special case of a more general problem whose problem instances take more general forms. For instance, Even and Tarjan [5] have extended the game Hex to general graphs and showed that the generalized Hex game, or the Shannon switching game on vertices, is PSPACE-complete, while the complexity of the more common version of Hex remained open. In this paper, we follow this approach to the group testing problem and demonstrate several completeness results on the generalized group testing problem.

We first introduce some terminologies about two-person query games. A *query history* is a set of queries together with their answers. The *solution space* associated with a query history H is the set of all objects in the domain which are *consistent* with the query history H. In other words, let $\mathrm{ANS}_x(y)$ denote the answer given by player A to the query y when x is the object to be identified. Then, the solution space associated with a query history $H = \{(y_1, a_1), \cdots, (y_m, a_m)\}$, where y_i's are queries and a_i's are corresponding answers, is the set $\{x \in \text{domain} | \mathrm{ANS}_x(y_i) = a_i \text{ for } i = 1, \cdots, m\}$. The *initial solution space* is simply the given domain. A shortest decision tree problem may thus be rephrased as the problem of using the minimum number of queries to reduce the solution space from the given domain to a singleton space.

We note that while the initial solution spaces often have simple structures, the solution spaces associated with arbitrary query histories may have complex structures. For example, it was pointed out in [18] that many researchers have conjectured that, for the sorting problem, the problem of determining the size of the solution space associated with a query history is #P-complete. The first two problems considered in this paper are concerned with the structure of the general solution spaces associated with given query histories. The first asks whether a given query history is consistent (or, whether the player A has been cheating), and the second asks what the size of the solution space associated with a given query history is.

CONSISTENCY PROBLEM. Given a domain D and a query history H, determine whether the query history H is consistent; i.e. whether the solution space associated with H is nonempty.

COUNTING PROBLEM. Given a domain D and a query history H, determine the size of the solution space associated with H.

Our third problem is concerned with the nonadaptive query games. In a nonadaptive query game, the player B must present a set of queries before he/she gets any answer from the player A [15]. Again, the goal here is to find a smallest set of queries which uniquely determines each object in the domain. The following problem asks a simpler recognition question of such a *determinant* set of queries.

DETERMINACY PROBLEM. Given a domain D and a set Q of queries, determine whether each set of answers to the queries in Q uniquely determines an object in the domain.

We will study the above questions in the context of the group testing problem. We consider several variations of the original group testing problem, derived from different domains and different oracles. In the following, for each set S, let $|S|$ denote the size of S; for each n and d, let $\mathscr{S}_n$ denote the set of all subsets of $\{1, \cdots, n\}$ and $\mathscr{S}_{n,d}$ the set of all sets S in $\mathscr{S}_n$ with $|S| = d$. For each pair of objects x and y, $\mathrm{ANS}_x(y)$ is the answer given by player A to query y when x is the object to be identified.

MODEL A_k $(k \geqq 1)$. Given a domain $\mathscr{S}_n$ and an answering function ANS_S (as the oracle) of the type

$$\mathrm{ANS}_S(T) = \begin{cases} i & \text{if } |S \cap T| = i < k, \\ k & \text{if } |S \cap T| \geqq k, \end{cases}$$

determine the set S.

MODEL A'_k $(k \geqq 1)$. Given a domain $\mathscr{S}_{n,d}$ and an answering function ANS_S of the same type as in Model A_k, determine the set S.

MODEL B. Given a domain $\mathscr{S}_n$ and an answering function ANS_S (as the oracle) of the type

$$\mathrm{ANS}_S(T) = \begin{cases} 0 & \text{if } S \cap T = \varnothing, \\ 1 & \text{if } S \cap T \neq \varnothing \text{ and } \bar{S} \cap T \neq \varnothing, \\ 2 & \text{if } \bar{S} \cap T = \varnothing, \end{cases}$$

where $\bar{S} = \{1, \cdots, n\} - S$, determine the set S.

MODEL B'. Given a domain $\mathscr{S}_{n,d}$ and an answering function ANS_S of the same type as in Model B, determine the set S.

MODEL C. Given a domain $\mathscr{S}_n$ and an answering function ANS_S of the type

$$\mathrm{ANS}_S(T) = |S \cap T|,$$

determine the set S.

MODEL C'. Given a domain $\mathscr{S}_{n,d}$ and an answering function ANS_S of the same type as in Model C, determine the set S.

We remark that Models A_1 and A'_1 are the original group testing problems [3], [7], [14], [23]–[26], [29]; Models A_k and A'_k, with $k > 1$, have been considered in [2], [9], [11], [13], [21], [27]; Models B and B' have been considered in [10]; and Models C and C' are a classical combinatorial search problem [1], [4], [19].

The main results of this paper may be summarized as follows.

THEOREM 1. (a) *The consistency problem for Model A_1 is polynomial time solvable.*

(b) *The consistency problems for all other models (i.e., for Models A_k, $k > 1$, for Models A'_k, $k \geqq 1$, and for Models B, B', C and C') are* NP*-complete.*

THEOREM 2. *The counting problems for all models (i.e., for Models A_k and A'_k, $k \geqq 1$, and for Models B, B', C and C') are* #P*-complete.*

THEOREM 3. (a) *The determinacy problems for all models are in co*-NP.

(b) *The determinacy problem for Model A_1 is polynomial time solvable.*

(c) *The determinacy problem for Models A_k, A'_k, $k \geqq 4$, and for Models B, B', C and C' are co*-NP-*complete.*

The question of whether the determinacy problems for Models A_k, $k = 2, 3$, and for Models A'_k, $k \leqq 3$, are co-NP-complete remains open.

2. Consistency problems. We first restate the consistency problems for the models of group testing problem defined in § 1. In the following, Consistency-X denotes the consistency problem for Model X, where $X \in \{A_k, A'_k, B, B', C, C' | k \geqq 1\}$.

CONSISTENCY-X. Given an integer n (or, two integers n and d) and a set $H = \{(T_j, a_j) | j = 1, \cdots, m\}$, with $T_j \in \mathcal{S}_n$, $a_j \in \{0, 1, \cdots, n\}$ for $j = 1, \cdots, m$, determine whether the set $C = \{S \in \mathcal{S}_n \text{ (or, } \mathcal{S}_{n,d}) | \text{ANS}_S(T_j) = a_j, j = 1, \cdots, m\}$ is nonempty.

It is interesting to observe the similarity between the group testing problem and the satisfiability problem (SAT) [8], where each query of the group testing problem may be regarded as a clause of variables for SAT. Therefore, our main tools for proving Theorems 1, 2 and 3 are variations of the satisfiability problem. For the proof of Theorem 1, we will use the following NP-complete problems.

VERTEX-COVER. Given a graph $G = (V, E)$ and an integer $k \leqq |V|$, determine whether there is a set $V' \subseteq V$ of size k such that each edge $e \in E$ is incident on some $v \in V'$.

ONE-IN-THREE-SAT. Given a set U of variables and a set $\mathcal{C}$ of clauses, with each $C \in \mathcal{C}$ containing exactly three variables from U, determine whether there is a truth assignment t on U such that each clause C in $\mathcal{C}$ contains exactly one TRUE variable.

NOT-ALL-EQUAL-SAT. Given U and $\mathcal{C}$ as in One-in-three-SAT, determine whether there is a truth assignment t on U such that each clause C in $\mathcal{C}$ contains at least one TRUE variable and at least one FALSE variable.

Remark. The original versions of One-in-three-SAT and Not-all-equal-SAT, as stated in [8], allow a clause C in $\mathcal{C}$ to contain both negated and nonnegated literals. The NP-completeness of our versions stated above can easily be proved from Schaefer's proof of the NP-completeness of the Generalized-SAT problem [22].

Now we apply these NP-complete problems to prove Theorem 1.

MODEL A_1. Let an instance $(n, H = \{(T_j, a_j) | j = 1, \cdots, m\})$ of Consistency-A_1 be given, where for each $j = 1, \cdots, m$, $T_j \in \mathcal{S}_n$ and $a_j \in \{0, 1\}$. Define

$$I = \{j | 1 \leqq j \leqq m, a_j = 0\},$$

and $J = \{j | 1 \leqq j \leqq m, a_j = 1\}$. Also let $X = \bigcup_{j \in I} T_j$ and $Y = \{1, \cdots, n\} - X$. Then, it is easy to check that

$$H \text{ is consistent iff for each } j \in J, T_j \cap Y \neq \varnothing.$$

This characterization of consistent query histories provides a simple polynomial-time algorithm for Consistency-A_1.

MODEL A_k, $k > 1$. We show that if $k > 1$, then One-in-three-SAT is polynomial-time reducible to Consistency-A_k.

Let an instance $(U, \mathcal{C})$ of One-in-three-SAT be given, where $U = \{x_1, \cdots, x_p\}$, $\mathcal{C} = \{C_1, \cdots, C_q\}$, $C_j \subseteq U$ and $|C_j| = 3$, for $j = 1, \cdots, q$. Define an instance

$$(n, H = \{(T_j, a_j) | j = 1, \cdots, m\})$$

of Consistency-A_k as follows:

$$n := p; \; m := q;$$

$$\text{for each } j = 1, \cdots, m, \text{ let } T_j := \{i | x_i \in C_j\} \text{ and } a_j := 1.$$

For each assignment t on U, let $S_t := \{i | t(x_i) = \text{TRUE}\}$. Then, the mapping from t to S_t is a natural one-to-one correspondence between the set of truth assignments on U and the set $\mathscr{S}_n$. Furthermore, a truth assignment t on U assigns exactly one TRUE variable to each clause in $\mathscr{C}$ if and only if $|S_t \cap T_j| = 1$ for all $j = 1, \cdots, m$. In other words, the instance $(U, \mathscr{C})$ has a solution (for the problem One-in-three-SAT) if and only if the instance (n, H) has a solution (for the problem Consistency-A_k). This completes the proof.

MODEL A'_1. We show that the problem Vertex-Cover is polynomial-time reducible to Consistency-A'_1.

Let (G, k) be a given instance of Vertex-Cover, where $G = (V, E)$ is a graph with vertex set $V = (v_1, \cdots, v_p\}$ and the edge set $E = \{e_1, \cdots, e_q\}$, and k is an integer less than or equal to p. Define an instance $(n, d, H = \{(T_j, a_j) | j = 1, \cdots, m\})$ of Consistency-A'_1 as follows.

$$n := p;\ m := q;\ d := k;$$

$$\text{for each } j = 1, \cdots, m, \text{ let } T_j := \{i | v_i \in e_j\} \text{ and } a_j := 1.$$

For each $V' \subseteq V$, define a set $S_{V'} \in \mathscr{S}_n$ by $S_{V'} = \{i | v_i \in V'\}$. Then, this is a one-to-one correspondence between subsets of V of size k and sets in $\mathscr{S}_{n,d}$. Furthermore, V' is a vertex cover of E if and only if $S_{V'} \cap T_j \neq \varnothing$ for all $j = 1, \cdots, m$. This shows that the mapping from (G, k) to (n, d, H) is a reduction from Vertex-Cover to Consistency-A'_1.

MODEL A'_k, $k > 1$. We show that if $k > 1$, then Consistency-A'_1 is polynomial-time reducible to Consistency-A'_k.

For a given instance $(n, d, H = \{(T_j, a_j) | j = 1, \cdots, m\})$ of Consistency-A'_1, define an instance $(n', d', H' = \{(T'_j, a'_j) | j = 1, \cdots, m\})$ of Consistency-A'_k as follows:

$$n' := n + k - 1;\ m' := m + k - 1;\ d' := d + k - 1;$$

$$\text{for each } j = 1, \cdots, m,$$

$$\text{if } a_j = 0 \text{ then let } T'_j := T_j \text{ and } a'_j := 0,$$

$$\text{if } a_j = 1 \text{ then let } T'_j := T_j \cup \{n+1, \cdots, n+k-1\} \text{ and } a'_j := k;$$

$$\text{for each } j = m+1, \cdots, m+k-1, \text{ let } T'_j := \{n+j-m\} \text{ and } a'_j := 1.$$

Assume that (n, d, H) is consistent for Model A'_1 and $S \in \mathscr{S}_{n,d}$ satisfies the condition that for all $j = 1, \cdots, m$, $S \cap T_j \neq \varnothing$ if and only if $a_j = 1$. Define

$$S' = S \cup \{n+1, \cdots, n+k-1\}.$$

Then, $S \in \mathscr{S}_{n',d'}$. Also, for all $j = 1, \cdots, m$,

$$\text{if } a_j = 0, \text{ then } |S' \cap T'_j| = |S \cap T_j| = 0 = a'_j, \quad \text{and}$$

$$\text{if } a_j = 1, \text{ then } |S' \cap T'_j| = |S \cap T_j| + (k-1) \geqq k = a'_j;$$

and, for all $j = m + 1, \cdots, m + k - 1$,

$$|S' \cap T'_j| = 1 = a'_j.$$

So, (n', d', H') is consistent for Model A'_k.

Conversely, if (n', d', H') is consistent for Model A'_k, then there is a set

$$S' \subseteq \{1, \cdots, n+k-1\}$$

such that $\text{ANS}_{S'}(T'_j) = a'_j$ for $j = 1, \cdots, m + k - 1$, where the answering function

$\mathrm{ANS}_{S'}$ is of the type of Model A'_k. Let $S = S' \cap \{1, \cdots, m\}$. We claim that $S \cap T_j = \varnothing$ if and only if $a_j = 0$ for all $j = 1, \cdots, m$.

First, if $a_j = 0$, then $T'_j = T_j$ and $a'_j = 0$. So, $S' \cap T'_j = \varnothing$ and hence $S \cap T_j = \varnothing$. Next, if $a_j = 1$, then $\mathrm{ANS}_{S'}(T'_j) = a'_j = k$ implies $|S' \cap T'_j| \geqq k$. Since

$$|S' \cap \{n+1, \cdots, n+k-1\}| \leqq k-1, \qquad |S \cap T_j| = |S' \cap T'_j \cap \{1, \cdots, n\}| \geqq 1.$$

This completes the proof for Model A'_k, $k > 1$.

MODEL B. We show that Not-all-equal-SAT is polynomial-time reducible to Consistency-B. The reduction is similar to the reduction from One-in-three-SAT to Consistency-A_2.

Let an instance $(U, \mathscr{C})$ of Not-all-equal-SAT be given, where $U = \{x_1, \cdots, x_p\}$, $\mathscr{C} = \{C_1, \cdots, C_q\}$, $C_j \subseteq U$ and $|C_j| = 3$, for $j = 1, \cdots, q$. Define an instance

$$(n, H = \{(T_j, a_j) | j = 1, \cdots, m\})$$

of Consistency-B as follows:

$$n := p; \; m := q;$$

$$\text{for each } j = 1, \cdots, m, \text{ let } T_j := \{i | x_i \in C_j\} \text{ and } a_j := 1.$$

Similarly to the reduction from One-in-three-SAT to Consistency-A_2, there is a natural one-to-one correspondence between the set of truth assignments on U and the set $\mathscr{S}_n$. Furthermore, a truth assignment t on U assigns at least one TRUE variable and at least one FALSE variable to each clause in $\mathscr{C}$ if and only if $1 \leqq |S_t \cap T_j| \leqq 2$ for all $j = 1, \cdots, m$, where S_t is the set in $\mathscr{S}_n$ corresponding to t. This shows that the mapping defined above is a reduction from Not-all-equal-SAT to Consistency-B.

MODEL B'. We show that Vertex-Cover is polynomial-time reducible to Consistency-B'.

Let (G, k) be a given instance of Vertex-Cover, where $G = (V, E)$ is a graph with the vertex set $V = \{v_1, \cdots, v_p\}$ and the edge set $E = \{e_1, \cdots, e_q\}$, and k is an integer less than or equal to p. Define an instance $(n, d, H = \{(T_j, a_j) | j = 1, \cdots, m\})$ of Consistency-B' as follows:

$$n := p + q; \; m := q; \; d := k;$$

$$\text{for each } j = 1, \cdots, m, \text{ assume that } e_j = \{v_{j_1}, v_{j_2}\}, \text{ and}$$

$$\text{let } T_j := \{j_1, j_2, p + j\} \text{ and } a_j := 1.$$

Let $V' \subseteq V$ be a vertex cover for G of size k. Then the set

$$S_{V'} = \{i | 1 \leqq i \leqq p, v_i \in V'\}$$

has the property $1 \leqq |S_{V'} \cap T_j| \leqq 2$ for all $j = 1, \cdots, m$. Also, $|S_{V'}| = d$. So, (n, d, H) is consistent.

Conversely, let $S \subseteq \{1, \cdots, n\}$, $|S| = d$, be a solution to the instance (n, d, H). Define $V' := \{v_i | i \in S, 1 \leqq i \leqq p\} \cup \{v_{j_1} | p + j \in S\}$. Then, $|V'| \leqq d = k$ because $|S| = d$. Also, V' is a vertex cover for G: for each $j = 1, \cdots, q$, if $p + j \notin S$ then j_1 or j_2 is in S and hence v_{j_1} or v_{j_2} is in V'; if $p + j \in S$, then $v_{j_1} \in V'$. This completes the proof.

MODEL C. The reduction from One-in-three-SAT to Consistency-A_2 is actually also a reduction from One-in-three-SAT to Consistency-C, because the output instances from the reduction always have $a_j = 1 < 2$.

MODEL C'. We show that Consistency-C is polynomial-time reducible to Consistency-C'.

Let an instance $(n, H = \{(T_j, a_j) | j = 1, \cdots, m\})$ of Consistency-C be given. Define an instance $(n', d', H' = \{(T'_j, a'_j) | j = 1, \cdots, m\})$ of Consistency-C' as follows:

$$n' := 2n;\ d' := n;\ m' := n + m;$$

$$\text{for } j = 1, \cdots, m, \text{ let } T'_j := T_j \text{ and } a'_j := a_j, \text{ and}$$

$$\text{for } j = m + 1, \cdots, m + n, \text{ let } T'_j := \{j - m, n + j - m\} \text{ and } a_j := 1.$$

If $S \in \mathscr{S}_n$ is consistent with H, define $S' = S \cup \{k + n | 1 \leqq k \leqq n, k \notin S\}$. Then, $S' \in \mathscr{S}_{n',d'}$, and

$$|S \cap T_j| = |S' \cap T'_j| \quad \text{for } j = 1, \cdots, m,$$

$$|S' \cap T'_j| = 1 \quad \text{for } j = m + 1, \cdots, m + n.$$

This shows that S' is consistent with H'.

Conversely, if $S' \in \mathscr{S}_{n',d'}$ is consistent with H', then $S = S' \cap \{1, \cdots, n\}$ satisfies the condition that for all $j = 1, \cdots, m$, $|S \cap T_j| = |S' \cap T'_j|$, because for all $j = 1, \cdots, m$, $T'_j \subseteq \{1, \cdots, n\}$ and so $|S' \cap T'_j| = |S' \cap T'_j \cap \{1, \cdots, n\}| = |S \cap T_j|$. This completes the proof.

3. Counting problems. We restate the counting problems for Model X, where $X \in \{A_k, A'_k, B, B', C, C' | k \geqq 1\}$.

COUNTING-X. Given an integer n (or, two integers n and d) and a set $H = \{(T_j, a_j) | j = 1, \cdots, m\}$, with $T_j \in \mathscr{S}_n$, $a_j \in \{0, 1, \cdots, n\}$ for $j = 1, \cdots, m$, determine the size of the set $C = \{S \in \mathscr{S}_n \text{ (or, } \mathscr{S}_{n,d}) | \mathrm{ANS}_S(T_j) = a_j, j = 1, \cdots, m\}$.

It is easy to see that for any model X, the problem Counting-X is in #P because the problem Consistency-X is in NP. (For the definitions of the class #P and #P-completeness, see [8] and [28].) In this section, we show that the counting problems for all models are #P-complete. We remark that this type of #P-completeness results has been conjectured for the sorting problem [18] and has been proved for a simplified Mastermind game [17]. The following #P-complete problems will be used in the proof of Theorem 2.

MONOTONE-#2SAT. Given a set U of variables and a set $\mathscr{C}$ of clauses, with each $C \in \mathscr{C}$ containing exactly two variables from U, determine the number of truth assignments t on U such that each clause C in $\mathscr{C}$ contains at least one TRUE variable.

ONE-IN-THREE-#SAT. Given $(U, \mathscr{C})$ as in One-in-three-SAT, determine the number of solutions to $(U, \mathscr{C})$.

NOT-ALL-EQUAL-#SAT. Given $(U, \mathscr{C})$ as in Not-all-equal-SAT, determine the number of solutions to $(U, \mathscr{C})$.

Monotone-#2SAT has been shown in [28] to be #P-complete. We first establish the #P-completeness of One-in-three-#SAT and Not-all-equal-#SAT. We note that a counting problem is not a decision problem and hence the polynomial-time many-one reductions are not necessarily applicable to them. Instead, the polynomial-time Turing reductions are usually used to prove the #P-completeness results, although the notion of many-one reductions preserving the number of solutions (or, *parsimonious reductions*) does provide a stronger definition of #P-completeness (cf. [8]). In this section, we refer to #P-completeness as the one with respect to the polynomial-time Turing reductions.

LEMMA 1. *One-in-three-*#SAT *is* #P-*complete.*

Proof. The fact that One-in-three-#SAT is in #P is clear. We show that Monotone-#2SAT is polynomial-time Turing reducible to One-in-three-#SAT.

Let an instance $(U, \mathscr{C})$ of Monotone-#2SAT be given, where $U = \{x_1, \cdots, x_p\}$, $\mathscr{C} = \{C_1, \cdots, C_q\}$ and for each $j = 1, \cdots, q$, $C_j \subseteq U$ and $|C_j| = 2$. Define an instance

$(V, \mathscr{D})$ of One-in-three-#SAT as follows:

$$V := U \cup \{u_j, v_j, w_j \mid j = 1, \cdots, q\} \cup \{y_1, y_2, y_3, z\};$$

for each $j = 1, \cdots, q$, assume that $C_j = \{x_{j_1}, x_{j_2}\}$ and let

$$C_{j,1} := \{x_{j_1}, u_j, y_1\}, C_{j,2} := \{x_{j_2}, v_j, y_1\}, C_{j,3} := \{u_j, v_j, w_j\};$$

let $D_1 := \{y_1, y_2, z\}$, $D_2 := \{y_2, y_3, z\}$, $D_3 := \{y_1, y_3, z\}$;

$$\mathscr{D} = \{C_{j,k} \mid j = 1, \cdots, q; k = 1, 2, 3\} \cup \{D_1, D_2, D_3\}.$$

Assume that t is a truth assignment on U such that for each $j = 1, \cdots, q$, there is a variable x_k in C_j with $t(x_k) =$ TRUE. Define a truth assignment t' on V as follows:

for each $i = 1, \cdots, p$, $t'(x_i) = t(x_i)$;
$t'(y_1) = t'(y_2) = t'(y_3) :=$ FALSE; $t'(z) :=$ TRUE;
for each $j = 1, \cdots, q$, assuming that $C_j = \{x_{j_1}, x_{j_2}\}$,
Case 1. if $t(x_{j_1}) =$ TRUE, $t(x_{j_2}) =$ FALSE
then $t'(u_j) = t'(w_j) :=$ FALSE and $t'(v_j) :=$ TRUE;
Case 2. if $t(x_{j_1}) =$ FALSE, $t(x_{j_2}) =$ TRUE
then $t'(v_j) = t'(w_j) :=$ FALSE and $t'(u_j) :=$ TRUE;
Case 3. if $t(x_{j_1}) = t(x_{j_2}) =$ TRUE
then $t'(u_j) = t'(v_j) :=$ FALSE and $t'(w_j) :=$ TRUE.

It is easy to check that t' assigns the value TRUE to exactly one variable in each clause in $\mathscr{D}$. Therefore, each solution t of $(U, \mathscr{C})$ is mapped to a solution t' of $(V, \mathscr{D})$, and the mapping is one-to-one.

Furthermore, we note that if t'' is a solution of $(V, \mathscr{D})$ then, to assign exactly one TRUE value to each of D_1, D_2 and D_3, t'' must assign TRUE to z and FALSE to y_1, y_2, y_3. Furthermore, for each $j = 1, \cdots, q$, t'' cannot assign the value TRUE to both u_j and v_j; this implies that one of x_{j_1} and x_{j_2} must be TRUE. Finally, for each $j = 1, \cdots, q$, if two solutions t_1 and t_2 of $(U, \mathscr{C})$ agree at x_{j_1} and x_{j_2} and $t_1(y_1) = t_2(y_1) =$ FALSE, then they must agree at u_j, v_j and w_j. The above observations show that the mapping defined above (from t to t') is a bijection between the solutions of $(U, \mathscr{C})$ and the solutions of $(V, \mathscr{D})$. This completes the proof. □

LEMMA 2. *Not-all-equal-*#SAT *is* #P-*complete.*

Proof. Again, it is clear that Not-all-equal-#SAT is in #P, and we show that Monotone-#2SAT is polynomial-time Turing reducible to Not-all-equal-#SAT.

Let an instance $(U, \mathscr{C})$ of Monotone-#2SAT be given, where $U = \{x_1, \cdots, x_p\}$, $\mathscr{C} = \{C_1, \cdots, C_q\}$ and for each $j = 1, \cdots, q$, $C_j \subseteq U$ and $|C_j| = 2$. Define an instance $(V, \mathscr{D})$ of Not-all-equal-#SAT as follows:

$$V := U \cup \{u_j, v_j \mid j = 1, \cdots, q\} \cup \{y_1, y_2, y_3, z\};$$

for each $j = 1, \cdots, q$, assume that $C_j = \{x_{j_1}, x_{j_2}\}$ and let

$$C_{j,1} := \{x_{j_1}, u_j, z\}, C_{j,2} := \{x_{j_2}, v_j, z\},$$

$$C_{j,3} := \{x_{j_1}, x_{j_2}, u_j\}, C_{j,4} := \{u_j, v_j, y_1\},$$

$$C_{j,5} := \{u_j, v_j, y_2\}, C_{j,6} := \{u_j, v_j, y_3\};$$

let $D := \{y_1, y_2, y_3\}$;

$$\mathscr{D} = \{C_{j,k} \mid j = 1, \cdots, q; k = 1, \cdots, 6\} \cup \{D\}.$$

We first note that if t' is a solution of $(V, \mathscr{D})$ (i.e., t' is a truth assignment on V such that t' assigns at least one TRUE value and at least one FALSE value to each clause in $\mathscr{D}$), then $t'(u_j) \neq t'(v_j)$ for $j = 1, \cdots, q$, because $t'(y_1)$, $t'(y_2)$ and $t'(y_3)$ cannot be all equal.

Now, assume that t is a solution of $(U, \mathscr{C})$. Define a truth assignment t' on $V - \{y_1, y_2, y_3\}$ as follows:

```
$t'(z) :=$ FALSE;
for each $i := 1, \cdots, p$, $t'(x_i) := t(x_i)$;
for each $j = 1, \cdots, q$, assuming that $C_j = \{x_{j_1}, x_{j_2}\}$,
    Case 1. if $t(x_{j_1}) =$ TRUE, $t(x_{j_2}) =$ FALSE
            then $t'(u_j) :=$ FALSE and $t'(v_j) :=$ TRUE;
    Case 2. if $t(x_{j_1}) =$ FALSE, $t(x_{j_2}) =$ TRUE
            then $t'(u_j) :=$ TRUE and $t'(v_j) :=$ FALSE;
    Case 3. if $t(x_{j_1}) = t(x_{j_2}) =$ TRUE
            then $t'(u_j) :=$ FALSE and $t'(v_j) :=$ TRUE.
```

We then extend t' into truth assignments on V such that $t'(y_1)$, $t'(y_2)$, $t'(y_3)$ are not all equal. There are six such extensions. It is obvious that each of these extensions is a solution of $(V, \mathscr{D})$. Next, for each of such extensions t'', define $\bar{t}''(w)$ to be the negation of $t''(w)$ for all $w \in V$. We get six more truth assignments which are solutions of $(V, \mathscr{D})$. (For the problem Not-all-equal-SAT, the negation of any solution is itself a solution.) We note that all these assignments are distinct. Furthermore, two distinct solutions t_1 and t_2 of $(U, \mathscr{C})$ define two disjoint sets of solutions of $(V, \mathscr{D})$. To see this, if a solution t''_1 of $(V, \mathscr{D})$ derived from t_1 is equal to a solution t''_2 of $(V, \mathscr{D})$ derived from t_2, then $t''_1(z) = t''_2(z)$. Hence, either $t_1 = t''_1|_U = t''_2|_U = t_2$ or $\bar{t}_1 = t''_1|_U = t''_2|_U = \bar{t}_2$, where $\bar{t}_1$ and $\bar{t}_2$ are the negations of t_1 and t_2, respectively. So, we get

$$12 \cdot (\text{\# of solutions of } (U, \mathscr{C})) \leqq \text{\# of solutions of } (V, \mathscr{D}).$$

Now, if t'' is a solution of $(U, \mathscr{C})$ then, as shown above, $t''(u_j) \neq t''(v_j)$ for all $j = 1, \cdots, q$. Assume that $t''(z) =$ FALSE. Then, to assign at least one TRUE value to both $C_{j,1}$ and $C_{j,2}$, at least one of $t''(x_{j_1})$ and $t''(x_{j_2})$ must be TRUE. Thus, $t''|_U$ is a solution of $(U, \mathscr{C})$, and t'' must be one of those 12 assignments defined by $t = t''|_U$. Similarly, if $t''(z) =$ TRUE, then $t = \bar{t}''|_U$ is a solution of $(U, \mathscr{C})$ and t'' is one of the 12 assignments defined by t. So, this shows that the number of solutions of $(V, \mathscr{D})$ is exactly 12 times the number of solutions of $(U, \mathscr{C})$. This completes the proof. □

With Lemmas 1 and 2, Theorem 2 is easy to prove. First, we show that for each model X, with $X \in \{A_k, B, C | k \geqq 1\}$, the problem Counting-$X$ is polynomial-time Turing reducible to Counting-X'.

LEMMA 3. *Let $X \in \{A_k, B, C | k \geqq 1\}$. Then, Counting-$X$ is polynomial-time Turing reducible to Counting-X'.*

Proof. Let (n, H) be an instance of Counting-X. Then, the number of sets S in $\mathscr{S}_n$ which are consistent with H is the sum of the number of sets S' in $\mathscr{S}_{n,d}$ which are consistent with H (with respect to the same type of answering functions) as d ranges over $\{0, \cdots, n\}$. □

MODEL A_1. We show that Monotone-#2SAT is a polynomial-time Turing reducible to Counting-A_1.

Let an instance $(U, \mathscr{C})$ of Monotone-#2SAT be given, where $U = \{x_1, \cdots, x_p\}$, $\mathscr{C} = \{C_1, \cdots, C_q\}$ and for each $j = 1, \cdots, q$, $C_j \subseteq U$ and $|C_j| = 2$. Define an instance $(n, H = \{(T_j, a_j) | j = 1, \cdots, m\})$ of Counting-A_1 as follows:

$$n := p; \ m := q;$$

$$\text{for each } j = 1, \cdots, m, \text{ let } T_j := \{i | x_i \in C_j\} \text{ and } a_j := 1.$$

Then, there is a natural one-to-one correspondence between truth assignments t on U and subsets S_t in $\mathcal{S}_n$, defined by $S_t = \{i | t(x_i) = \text{TRUE}\}$. This mapping also preserves the solutions of the two instances $(U, \mathscr{C})$ and (n, H). Thus, the number of solutions of these two instances are equal. This completes the proof.

MODEL A_k, $k > 1$ AND MODEL C. In § 2, it is proved that if $k > 1$ then One-in-three-SAT is polynomial-time (many-one) reducible to Consistency-A_k (and to Consistency-C). A close inspection of the reduction shows that the reduction actually preserves the number of solutions of the two problems. Thus, it also serves as a reduction from One-in-three-#SAT to Counting-A_k (and to Counting-C).

MODEL B. The polynomial-time (many-one) reduction from Not-all-equal-SAT to Consistency-B, as proved in § 2, also preserves the number of solutions. Thus, it also serves as a reduction from Not-all-equal-#SAT to Counting-B.

MODELS A'_k, $k \geqq 1$, MODEL B' AND MODEL C'. The #P-completeness of Counting-X', for $X \in \{A_k, B, C | k \geqq 1\}$, is established through Lemma 3 and the #P-completeness of Counting-X.

4. Determinacy problems. We restate the determinacy problems for Model X, where $X \in \{A_k, A'_k, B, B', C, C' | k \geqq 1\}$.

DETERMINACY-X. Given an integer n (or, two integers n and d) and a set $Q = \{T_j | j = 1, \cdots, m\}$, with $T_j \in \mathcal{S}_n$ for $j = 1, \cdots, m$, determine whether, for any two sets S_1, S_2 in $\mathcal{S}_n$ (or, in $\mathcal{S}_{n,d}$), $S_1 \neq S_2$ implies $\text{ANS}_{S_1}(T_j) \neq \text{ANS}_{S_2}(T_j)$ for some $j = 1, \cdots, m$.

We will call a set Q of queries *determinant* for Model X (with respect to size n) if the above problem Determinacy-X has an affirmative answer for input (n, Q). It is easy to see that for any model X, the problem Determinacy-X is in co-NP. We show, in this section, that most of them are actually co-NP-complete. Our main tools are the NP-complete problems One-in-three-SAT and Not-all-equal-SAT. Their precise definitions were given in § 2.

MODEL A_1. We give, in the following, a simple characterization of determinant sets Q of queries for Model A_1. This characterization provides a polynomial-time algorithm for Determinacy-A_1.

LEMMA 4. *A set Q is determinant for Model A_1 with respect to size n if and only if for every $i = 1, \cdots, n$, the singleton set $\{i\}$ is in Q.*

Proof. The backward direction is obvious, because the set $\{i\}$ distinguishes between two sets S_1 and S_2 whenever $i \in S_1 - S_2$.

For the forward direction, we consider two sets $S_1 = \{1, \cdots, n\}$ and $S_2 = S_1 - \{i\}$. Then, the only set T that can distinguish between S_1 and S_2 is $T = \{i\}$ so that $\text{ANS}_{S_1}(T) = 1$ and $\text{ANS}_{S_2}(T) = 0$. □

MODEL B. We show that Not-all-equal-SAT is polynomial-time reducible to the complement of Determinacy-B, and hence Determinacy-B is co-NP-complete.

Let an instance $(U, \mathscr{C})$ of Not-all-equal-SAT be given, where $U = \{x_1, \cdots, x_p\}$, $\mathscr{C} = \{C_1, \cdots, C_q\}$ and for each $j = 1, \cdots, q$, $C_j \subseteq U$ and $|C_j| = 3$. Define an instance

(n, Q) of Determinacy-B as follows:

$$\begin{aligned}&n := p;\\&\text{for each } j = 1, \cdots, q, \text{ let } T_{j,0} := \{i \mid x_i \in C_j\}, \text{ and}\\&\quad\text{for each } k = 1, \cdots, p, \text{ let } T_{j,k} := T_{j,0} \cup \{k\};\\&\text{let } Q = \{T_{j,k} \mid j = 1, \cdots, q; k = 0, \cdots, p\}.\end{aligned}$$

(Note that for each j, there are exactly $(p - 2)$ $T_{j,k}$'s; however, the total number of $T_{j,k}$'s in Q varies, depending on the set $\mathscr{C}$.)

Assume that t is a truth assignment on U such that for every $j = 1, \cdots, q$, t does not assign equal values to all three variables in C_j. Define $S_1 = \{i \mid t(x_i) = \text{TRUE}\}$ and $S_2 = \{1, \cdots, n\} - S_1$. Then, for each $j = 1, \cdots, q$, $S_1 \cap T_{j,0} \neq \varnothing$, and $S_2 \cap T_{j,0} \neq \varnothing$. This implies that for all $j = 1, \cdots, q$ and for all $k = 0, \cdots, n$, $\text{ANS}_{S_1}(T_{j,k}) = \text{ANS}_{S_2}(T_{j,k}) = 1$. So, Q is not determinant for Model B.

Conversely, assume that Q is not determinant and there are two sets $S_1, S_2 \subseteq \{1, \cdots, n\}$ such that $S_1 \neq S_2$ and $\text{ANS}_{S_1}(T_{j,k}) = \text{ANS}_{S_2}(T_{j,k})$ for all $j = 1, \cdots, q$, and $k = 0, \cdots, p$. Then, we claim that $\text{ANS}_{S_1}(T_{j,0})$ must be equal to 1 for all $j = 1, \cdots, q$.

Suppose, otherwise, that for some j, $\text{ANS}_{S_1}(T_{j,0}) = 0$ or 2. If $\text{ANS}_{S_1}(T_{j,0}) = 0$, then $T_{j,0} \cap S_1 = T_{j,0} \cap S_2 = \varnothing$. This implies that for any $k = 1, \cdots, p$,

$$\begin{aligned}x_k \in S_1 &\Leftrightarrow T_{j,k} \cap S_1 \neq \varnothing \Leftrightarrow \text{ANS}_{S_1}(T_{j,k}) = 1\\&\Leftrightarrow \text{ANS}_{S_2}(T_{j,k}) = 1 \Leftrightarrow T_{j,k} \cap S_2 \neq \varnothing \Leftrightarrow x_k \in S_2;\end{aligned}$$

or, $S_1 = S_2$. Similarly, if $\text{ANS}_{S_1}(T_{j,0}) = 2$, then $T_{j,0} \subseteq S_1$ and $T_{j,0} \subseteq S_2$. So, for any $k = 1, \cdots, p$,

$$\begin{aligned}x_k \in S_1 &\Leftrightarrow T_{j,k} \subseteq S_1 \Leftrightarrow \text{ANS}_{S_1}(T_{j,k}) = 2\\&\Leftrightarrow \text{ANS}_{S_2}(T_{j,k}) = 2 \Leftrightarrow T_{j,k} \subseteq S_2 \Leftrightarrow x_k \in S_2;\end{aligned}$$

or, $S_1 = S_2$. Both cases lead to contradictions. So the claim is proven.

Now, define a truth assignment t on U by $t(x_i) = \text{TRUE}$ if and only if $i \in S_1$. The claim that $\text{ANS}_{S_1}(T_{j,0}) = 1$, for all $j = 1, \cdots, q$, implies that t assigns at least one TRUE value and at least one FALSE value to each C_j in $\mathscr{C}$. This completes the proof.

MODEL B'. We show that Determinacy-B is polynomial-time reducible to Determinacy-B'.

Let an instance (p, Q) of Determinacy-B be given such that $Q = \{T_j \mid j = 1, \cdots, q\}$ and each T_j is in $\mathscr{S}_p$. Define an instance of Determinacy-B' as follows:

$$\begin{aligned}&n := 2p;\ m := 2q;\ d := p;\\&\text{for each } j = 1, \cdots, q, \text{ let } W_j := \{i + p \mid i \in T_j\};\\&\text{let } Q' := \{T_j, W_j \mid j = 1, \cdots, q\}.\end{aligned}$$

If Q is not determinant for Model B, then there are $S_1, S_2 \in \mathscr{S}_p$ such that $S_1 \neq S_2$ and for each $j = 1, \cdots, q$, $\text{ANS}_{S_1}(T_j) = \text{ANS}_{S_2}(T_j)$. Define $S_3 := S_1 \cup \{i + p \mid i \notin S_1\}$ and

$S_4 := S_2 \cup \{i + p | i \notin S_2\}$. Then, $S_3 \neq S_4$ and $|S_3| = |S_4| = p$. Furthermore, for each $j = 1, \cdots, q$.

$$\mathrm{ANS}_{S_3}(T_j) = \mathrm{ANS}_{S_1}(T_j) = \mathrm{ANS}_{S_2}(T_j) = \mathrm{ANS}_{S_4}(T_j), \quad \text{and}$$

$$\mathrm{ANS}_{S_3}(W_j) = 2 - \mathrm{ANS}_{S_1}(T_j) = 2 - \mathrm{ANS}_{S_2}(T_j) = \mathrm{ANS}_{S_4}(W_j).$$

So, Q' is not determinant for Model B'.

Conversely, if Q' is not determinant for Model B', then there exist $S_3, S_4 \in \mathcal{S}_n$ such that $S_3 \neq S_4$ and for each $j = 1, \cdots, q$, $\mathrm{ANS}_{S_3}(T_j) = \mathrm{ANS}_{S_4}(T_j)$ and $\mathrm{ANS}_{S_3}(W_j) = \mathrm{ANS}_{S_4}(W_j)$. Since $S_3 \neq S_4$, either $S_3 \cap \{1, \cdots, p\} \neq S_4 \cap \{1, \cdots, p\}$ or $S_3 \cap \{p+1, \cdots, 2p\} \neq S_4 \cap \{p+1, \cdots, 2p\}$. In the former case, we define $S_1 := S_3 \cap \{1, \cdots, p\}$ and $S_2 := S_4 \cap \{1, \cdots, p\}$; and, in the latter case, $S_1 := \{i | i + p \in S_3\}$ and $S_2 := \{i | i + p \in S_4\}$. Then, $S_1 \neq S_2$ but for each $j = 1, \cdots, q$, $\mathrm{ANS}_{S_1}(T_j) = \mathrm{ANS}_{S_2}(T_j)$. So, Q is not determinant for Model B.

MODEL C. We first simplify the problem.

LEMMA 5. *Let $Q = \{T_j | j = 1, \cdots, m\}$ be given such that $T_j \in \mathcal{S}_n$ for all $j = 1, \cdots, m$. Then, Q is not determinant for Model C, with respect to size n, if and only if there exist $S_1, S_2 \in \mathcal{S}_n$ such that $S_1 \cup S_2 \neq \varnothing$, $S_1 \cap S_2 = \varnothing$ and for each $j = 1, \cdots, m$, $|S_1 \cap T_j| = |S_2 \cap T_j|$.*

Proof. The backward direction is obvious. For the forward direction, we note that if S'_1 and S'_2 are two sets in $\mathcal{S}_n$ such that $S'_1 \neq S'_2$ and for each $j = 1, \cdots, m$, $|S'_1 \cap T_j| = |S'_2 \cap T_j|$. Then, the sets $S_1 = S'_1 - S'_2$ and $S_2 = S'_2 - S'_1$ satisfy the required condition. □

We now show that One-in-three-SAT is polynomial-time reducible to the complement of Determinacy-C.

Let $(U, \mathcal{C})$ be a given instance of One-in-three-SAT such that $U = \{x_1, \cdots, x_p\}$, $\mathcal{C} = \{C_1, \cdots, C_q\}$ and for every $j = 1, \cdots, q$, $C_j \subseteq U$ and $|C_j| = 3$. Without loss of generality, we assume that every x_i in U occurs in some C_j in $\mathcal{C}$. Define an instance (n, Q) of Determinacy-C as follows:

$n := p + 9q + 1$; $m := 10q$;

for convenience, for each $j = 1, \cdots, q$, and $k = 1, 2, 3$, let

$$u(j,k) := p + 9(j-1) + k,$$

$$v(j,k) := p + 9(j-1) + k + 3,$$

$$w(j,k) := p + 9(j-1) + k + 6;$$

also let $y := p + 9q + 1$;

for each $j = 1, \cdots, q$, assume that $C_j = \{x_{j_1}, x_{j_2}, x_{j_3}\}$ (with $j_1 < j_2 < j_3$), and define

$$T_{j,0} := \{j_1, j_2, j_3, y\},$$

$$T_{j,1} := \{j_2, j_3, u(j,1), v(j,1)\},$$

$$T_{j,2} := \{j_1, j_3, u(j,2), v(j,2)\},$$

$$T_{j,3} := \{j_1, j_2, u(j,3), v(j,3)\},$$

for each $j = 1, \cdots, q$, and each $k = 1, 2, 3$, define

$$U_{j,k} := \{u(j,k), w(j,k)\} \quad \text{and} \quad V_{j,k} := \{v(j,k), w(j,k)\};$$

let $Q := \{T_{j,h}, U_{j,k}, V_{j,k} | j = 1, \cdots, q; h = 0, \cdots, 3; k = 1, 2, 3\}$.

Assume that t is a truth assignment on U such that for each $C_j \in \mathscr{C}$, t assigns exactly one TRUE value to the variables in C_j. Define two sets $S_1, S_2 \in \mathscr{S}_n$ as follows:

$$S_1 := \{i \mid 1 \leqq i \leqq p, t(x_i) = \text{TRUE}\} \cup \{u(j,k), v(j,k) \mid \text{the } k\text{th variable } x_{j_k} \text{ in } C_j \text{ has } t(x_{j_k}) = \text{TRUE}\} \cup \{y\},$$

$$S_2 := \{i \mid 1 \leqq i \leqq p, t(x_i) = \text{FALSE}\} \cup \{w(j,k) \mid \text{the } k\text{th variable } x_{j_k} \text{ in } C_j \text{ has } t(x_{j_k}) = \text{TRUE}\}.$$

Obviously, $S_1 \cup S_2 \neq \varnothing$, $S_1 \cap S_2 = \varnothing$. We claim that for all $R \in Q$, $|S_1 \cap R| = |S_2 \cap R|$. For each $j = 1, \cdots, q$, we check the following:

(i) $|S_1 \cap T_{j,0}| = |S_2 \cap T_{j,0}|$: Among $\{j_1, j_2, j_3\}$, one is in S_1 and two are in S_2; and y is in S_1.

(ii) For $k = 1, 2, 3$, $|S_1 \cap T_{j,k}| = |S_2 \cap T_{j,k}|$: If $t(x_{j_1}) = \text{TRUE}$ and $t(x_{j_2}) = t(x_{j_3}) = \text{FALSE}$, then $u(j, 1)$, $v(j, 1)$ are in S_1. So, $S_1 \cap T_{j,1} = \{u(j, 1), v(j, 1)\}$ and $S_2 \cap T_{j,1} = \{j_1, j_2\}$; and $S_1 \cap T_{j,2} = S_1 \cap T_{j,3} = \{j_1\}$, $S_2 \cap T_{j,2} = \{j_3\}$ and $S_2 \cap T_{j,3} = \{j_2\}$. The other two cases are similar.

(iii) For $k = 1, 2, 3$, $|S_1 \cap U_{j,k}| = |S_2 \cap U_{j,k}|$ and $|S_1 \cap V_{j,k}| = |S_2 \cap V_{j,k}|$: From the definitions of S_1 and S_2, for any $j = 1, \cdots, q$ and $k = 1, 2, 3$, $u(j, k) \in S_1 \Leftrightarrow w(j, k) \in S_2 \Leftrightarrow v(j, k) \in S_1$.

Conversely, assume that Q is not determinant for Model C. Then, by Lemma 5, there exist $S_1, S_2 \in \mathscr{S}_n$ such that $S_1 \cup S_2 \neq \varnothing$, $S_1 \cap S_2 = \varnothing$ and for all $R \in Q$, $|S_1 \cap R| = |S_2 \cap R|$. First note the following fact:

(iv) For all $j = 1, \cdots, q$ and $k = 1, 2, 3$,

$$u(j,k) \in S_1 \Leftrightarrow w(j,k) \in S_2 \Leftrightarrow v(j,k) \in S_1, \quad \text{and}$$

$$u(j,k) \in S_2 \Leftrightarrow w(j,k) \in S_1 \Leftrightarrow v(j,k) \in S_2.$$

Next, we claim the following properties (v) and (vi).

(v) For any $j = 1, \cdots, q$, $|S_1 \cap T_{j,0}| \neq 1$.

Proof of (v). Assume otherwise that $|S_1 \cap T_{j,0}| = 1$. Then $|S_2 \cap T_{j,0}| = 1$. The following case analysis shows that this leads to a contradiction.

Case 1. $S_1 \cap T_{j,0} = \{j_1\}$, $S_2 \cap T_{j,0} = \{j_2\}$. Then, $j_2, j_3 \notin S_1$ and $j_1, j_3 \notin S_2$. So, $S_1 \cap T_{j,1} = S_1 \cap \{u(j, 1), v(j, 1)\}$ and $S_2 \cap T_{j,1} = \{j_2\} \cup (S_2 \cap \{u(j, 1), v(j, 1)\})$. By fact (iv) and the fact that $S_1 \cap S_2 = \varnothing$, we can see that $|S_1 \cap T_{j,1}| \neq |S_2 \cap T_{j,1}|$. This is a contradiction.

Case 2. $S_1 \cap T_{j,0} = \{j_1\}$, $S_2 \cap T_{j,0} = \{y\}$. Then, $j_2, j_3 \notin S_1$ and $j_1, j_2, j_3 \notin S_2$. So, $S_1 \cap T_{j,2} = \{j_1\} \cup (S_1 \cap \{u(j, 2), v(j, 2)\})$ and $S_2 \cap T_{j,2} = S_2 \cap \{u(j, 2), v(j, 2)\}$. Again, a contradiction.

Other cases. All other cases are symmetric to either Case 1 or Case 2.

(vi) $\{1, \cdots, p\} \subseteq S_1 \cup S_2$.

Proof of (vi). Assume otherwise that there is an i, $1 \leqq i \leqq p$, such that $i \notin S_1 \cup S_2$. Assume, without loss of generality, that x_i occurs as the first variable in C_j for some $j = 1, \cdots, q$; i.e., $x_i = x_{j_1}$.

Since $j_1 \notin S_1 \cup S_2$ and $S_1 \cap S_2 = \varnothing$, $|S_1 \cap T_{j,0}| = |S_2 \cap T_{j,0}| \leqq 1$. By claim (v), $S_1 \cap T_{j,0} = S_2 \cap T_{j,0} = \varnothing$. So, $y \notin S_1 \cup S_2$. However, this implies that for all $h = 1, \cdots, q$, $|S_1 \cap T_{h,0}| = |S_2 \cap T_{h,0}| \leqq 1$, and hence, by claim (v), $S_1 \cap T_{h,0} = S_2 \cap T_{h,0} = \varnothing$. This implies that $\{1, 2, \cdots, p\} \cap (S_1 \cup S_2) = \varnothing$.

In addition, fact (iv) shows that for any $h = 1, \cdots, q$ and $k = 1, 2, 3$, $|S_1 \cap T_{h,k}|$ is either 0 or 2. Since $|S_1 \cap T_{h,k}| = 2$ would imply that $|S_2 \cap T_{h,k}| = 0$ and make

$$|S_1 \cap T_{h,k}| \neq |S_2 \cap T_{h,k}|,$$

we must have $S_1 \cap T_{h,k} = \emptyset$. As a consequence, $S_1 = S_2 = \emptyset$. This is a contradiction, and so (vi) is proven.

Now we complete the proof of the reduction. Since $\{1, \cdots, p\} \subseteq S_1 \cup S_2$, y must be in $S_1 \cup S_2$. Assume, without loss of generality, that $y \in S_1$. Define a truth assignment t on U by $t(x_i) = \text{TRUE}$ if and only if $i \in S_1$. Then, for each $j = 1, \cdots, q$, $|S_1 \cap T_{j,0}| = |S_2 \cap T_{j,0}|$ implies that $|S_1 \cap T_{j,0}| = 2$. Since $y \in S_1$, $|S_1 \cap \{j_1, j_2, j_3\}| = 1$. That is, there is exactly one $k \in \{1, 2, 3\}$ such that $t(x_{j_k}) = \text{TRUE}$. This completes the proof for Model C.

MODEL A_k, $k \geqq 4$. Assume that $k \geqq 4$ and that Q is a set of queries each of size $\leqq 4$. Then, Q is determinant for Model C if and only if Q is determinant for Model A_k, because the answering functions for both models behave exactly the same on queries of size $\leqq 4$. In the above, for the problem Determinacy-C, we have actually shown a reduction from One-in-three-SAT to the complement of the following special case of Determinacy-C.

DETERMINACY-$C4$. Given an integer n and a set Q of queries each of size $\leqq 4$, determine whether Q is determinant for Model C with respect to size n.

From the above discussion, this problem is also a special case for Model A_k. So, it also proves that Determinacy-A_k is co-NP-complete.

MODEL C' AND MODEL A'_k, $k \geqq 4$. We can show that Determinacy-$C4$ is polynomial-time reducible to Determinacy-C' and Determinacy-A'_k, for $k \geqq 4$. The reductions are similar to the reduction from Determinacy-B to Determinacy-B'. The key point is that for the answering functions for Model C and Model A_k, $k \geqq 4$, the following property holds for all T of size $\leqq 4$:

$$\text{ANS}_S(T) = |T| - \text{ANS}_{\bar{S}}(T),$$

where $\bar{S} = \{1, \cdots, n\} - S$. This property allows us to carry out the reductions as in the case for Determinacy-B'. We omit the details. (Note that the above property holds for queries T of any size if we only consider Model C. However, for Model A_k, $k \geqq 1$, it only holds for queries T of size $\leqq k$.)

5. Discussion. In the last three sections, we have demonstrated several NP-hardness results on problems related to group testing. The NP-completeness of the consistency problems and the #P-completeness of the counting problems show that the solution spaces associated with arbitrary query histories have complex structures. The co-NP-completeness of the determinacy problems shows that the recognition version of the nonadaptive group testing problems is intractable. It is interesting to compare this problem with the problem of finding a minimal determinant set for Model C, for which a polynomial-time almost-optimal algorithm has been found by Cantor and Mills [1] and Linström [19].

While the complexity for the above three problems has been characterized precisely for most models considered, we have left many more questions open. To name the most important ones, we consider the following two problems concerned with the minimization of the heights of decision trees in the generalized form.

MINIMUM TEST PROBLEM. Given a domain D, a query history H and an integer k, determine whether there is a decision tree of height $\leqq k$ such that each path of the decision tree uniquely determines an object in the solution space associated with the query history H.

MINIMUM NONADAPTIVE TEST PROBLEM. Given a domain D, a query history H and an integer k, determine whether there is a set Q of k queries such that each set of answers to the queries in Q uniquely determines an object in the solution space associated with the query history H.

In the above, the minimum test problem is the generalization of the basic shortest decision tree problem we discussed in § 1, and the minimum nonadaptive test problem is the corresponding problem for the nonadaptive case. It is not hard to see that for models considered in this paper, the minimum nonadaptive test problems are in Σ_2^P, and the minimum test problems are in PSPACE, where Σ_2^P is the class of languages recognized by nondeterministic oracle Turing machines in polynomial time relative to oracle sets in NP [8], and PSPACE is the class of languages recognized by deterministic Turing machines in polynomial space [8]. Furthermore, the proofs of the NP-completeness of the consistency problems can easily be modified to show the NP-hardness of the minimum nonadaptive test problems and the minimum test problems for the same models. In view of the difficulty of getting optimal algorithms for these problems even for simple initial solution spaces and the complex structure of general solution spaces, we conjecture that the minimum nonadaptive test problems for most models are Σ_2^P-complete and the minimum test problem for most models are PSPACE-complete.

Other interesting questions include the following:

(1) Instead of the query history, we may use different representations for a solution space, for example, by listing its elements explicitly. What are the effects of these different representations of solution spaces on the computational complexity of the questions considered here?

(2) Do these NP-hardness results hold for the group testing problems with respect to the average-case complexity?

(3) Can we prove completeness results for other searching problems which involve the minimization of the heights of decision trees?

Acknowledgment. The authors thank Professor Ronald Book for his suggestions and encouragement.

REFERENCES

[1] D. G. CANTOR AND W. H. MILLS, *Determination of a subset from certain combinatorial properties*, Canad. J. Math., 18 (1966), pp. 42–48.

[2] X. M. CHANG AND F. K. HWANG, *The minimax number of calls for finite population multiple-access channel*, manuscript.

[3] R. DORFMAN, *The detection of defective members in large population*, Ann. Math. Statist., 14 (1943), pp. 4436–4440.

[4] P. ERDÖS AND A. RENYI, *On two problems of information theory*, Publ. Hung. Acad. Sci., 8 (1963), pp. 241–254.

[5] S. EVEN AND R. E. TARJAN, *A combinatorial problem which is complete in polynomial space*, J. Assoc. Comput. Mach., 23 (1976), pp. 710–719.

[6] S. FORTUNE, *A note on sparse complete sets*, SIAM J. Comput., 8 (1979), pp. 431–433.

[7] M. R. GAREY, *Optimal binary identification procedures*, SIAM J. Appl. Math., 23 (1972), pp. 173–186.

[8] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[9] E. P. GUDJOHNSEN, D. TOWSLEY AND J. K. WOLF, *On adaptive polling techniques for computer communication networks*, ICC 80 Conference Record 1, Seattle, WA, 1980, pp. 13.3.1–13.3.5.

[10] F. K. HWANG, *Three versions of a group testing game*, this Journal, 5 (1984), pp. 145–153.

[11] ———, *When is group testing querying profitable in communication networks?*, manuscript.

[12] ———, Private communications.

[13] F. K. HWANG AND X. M. CHANG, *Cutoff points for roll call protocals in multiple access system*, IEEE Trans. Inform. Theory, 33 (1987), pp. 577–581.

[14] F. K. HWANG, T. T. SONG AND D. Z. DU, *Hypergeometric group testing and generalized hypergeometric group testing problems*, this Journal, 2 (1981), pp. 426–428.

[15] F. K. HWANG AND V. T. SOS, *Non-adaptive hypergeometric group testing*, Studia Sci. Math. Hungar., to appear.

[16] D. E. KNUTH, *The Art of Computer Programming*, Vol. 3, *Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.

[17] K. KO AND S. TENG, *On the number of queries necessary to identify a permutation*, J. Algorithms, 7 (1986), pp. 449–462.

[18] N. LINIAL, *The information-theoretic bound is good for merging*, SIAM J. Comput., 13 (1984), pp. 795–801.

[19] B. LINSTRÖM, *On a combinatorial problem in number theory*, Canad. Math. Bull., 8 (1965), pp. 477–490.

[20] G. K. MANACHER, *The Ford–Johnson sorting algorithm is not optimal*, J. Assoc. Comput. Mach., 26 (1979), pp. 441–456.

[21] N. MEHRAVARI, *Generalized binary binomial group testing*, this Journal, 7 (1986), pp. 159–166.

[22] T. J. SCHAEFER, *The complexity of satisfiability problems*, Proc. 10th Annual ACM Symposium on Theory of Computing, 1978, pp. 216–226.

[23] M. SOBEL, *Binomial and hypergeometric group-testing*, Studia Sci. Math. Hungar., 2 (1968), pp. 19–42.

[24] M. SOBEL AND P. A. GROLL, *Group-testing to eliminate efficiently all defectives in a binomial sample*, Bell System Tech. J., 38 (1959), pp. 1179–1252.

[25] A. STERRETT, *On the decision of defective members of large population*, Ann. Math. Statist., 28 (1957), pp. 1033–1036.

[26] J. THOMAS, B. S. PASTERNACK, S. J. VACIRCA AND D. L. THOMPSON, *Application of group testing procedures in radiological health*, Health Physics, 25 (1973), pp. 259–266.

[27] B. S. TSYBAKOV, V. A. MIKHAILOV AND N. B. LIKHANOV, *Bounds for packet transmission rate in a random-multiple-access system*, Problemy Peredachi Informatsii, 19 (1983), pp. 61–81. (In Russian.)

[28] L. G. VALIANT, *The complexity of computing the permanent*, Theoret. Comput. Sci., 9 (1979), pp. 189–201.

[29] G. S. WATSON, *A study of the group screening method*, Technometrics, 3 (1961), pp. 371–388.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

021

A RECURSIVE HADAMARD TRANSFORM OPTIMAL SOFT DECISION DECODING ALGORITHM*

YAIR BE'ERY† AND JAKOV SNYDERS†

Abstract. A recursive soft decision maximum likelihood Hadamard transform decoding rule for binary code is derived. This algorithm, with computational complexity that varies inversely with the code rate for a fixed code length, is efficiently applicable for decoding convolutional codes and high rate block codes. An even more significant reduction in decoder complexity is obtained when the algorithm is applied for decoding product codes and concatenated codes. This algorithm achieves the computational efficiency of the Viterbi algorithm. In addition, its structural regularity simplifies the VLSI implementation of decoders.

Key words. maximum likelihood decoding, soft decision, Hadamard transform, error correcting codes

AMS(MOS) subject classifications. primary 94B35; secondary 94B05, 94B10

1. Introduction. Let $\mathscr{C}$ be an (n, k) binary linear block code of length n and dimension k, and let $\mathbf{g}_i$, $i = 0, 1, \cdots, n-1$, be the columns of the generator matrix G of $\mathscr{C}$ that represent the encoding, i.e., a message $\mathbf{s} \in \mathrm{GF}(2)^k$ is mapped onto $\mathbf{c} \in \mathscr{C}$ according to $\mathbf{c} = \mathbf{s}G$. Denote the real line by R. Assume that codewords with equal probability 2^{-k} are transmitted through a memoryless channel characterized by transition probability densities $f_j(v) = f(v|j)$ where $v \in R$ and $j \in \mathrm{GF}(2)$. Let $\mathbf{v} = (v_0, v_1, \cdots, v_{n-1})$, $v_i \in R$ be the word observed at the output of the channel. Our aim is to find a codeword $\mathbf{c} = (c_0, c_1, \cdots, c_{n-1}) \in \mathscr{C}$ that maximizes the probability density $f(\mathbf{v}|\mathbf{c}) = \prod_{i=0}^{n-1} f(v_i|c_i)$. As is shown in [1], we may instead seek $\mathbf{s} \in \mathrm{GF}(2)^k$ that maximizes $M(\cdot)$ given by

$$M(\mathbf{s}) = \sum_{i=0}^{n-1} (-1)^{\langle \mathbf{s}, \mathbf{g}_i \rangle} \mu(v_i) \tag{1}$$

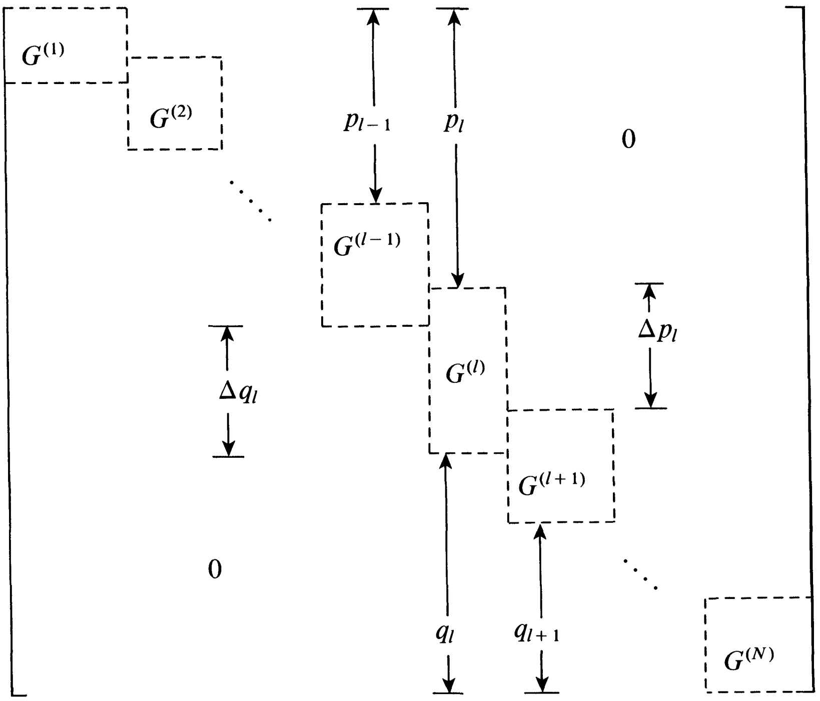
where $\langle \cdot, \cdot \rangle$ stands for inner product over GF(2) and $\mu(v) = \log(f_0(v)/f_1(v))$. (In the case of a discrete output-alphabet, that replaces R, the same expression results except that the log-likelihood ratio $\mu(\cdot)$ is defined in terms of transition probabilities.) This procedure requires $(n-1)2^k + (2^k - 1) = n2^k - 1$ addition-equivalent operations, where $2^k - 1$ accounts for the complexity of maximization. By slight modification (1) becomes a Hadamard transform, thereby allowing reduction of computational complexity with the aid of a fast algorithm [1].

* Received by the editors June 2, 1986; accepted for publication July 29, 1987. The research reported here was performed in part at McGill University, Montréal, Québec 43A2K6, Canada and at Rensselaer Polytechnic Institute, Troy, New York 12181, where the second author was on leave.

† Department of Electronic Communications, Control and Computer Systems, Tel-Aviv University, Tel-Aviv 69978, Israel.

In this paper we assume that $\mathscr{C}$ is generated by a $k \times n$ matrix G which possesses the following double-echelon structure:

(2) $G=$



where $N > 1$ (actually, $N >> 1$ in most applications). Above and beneath each submatrix $G^{(l)}$ of dimension $k_l \times n_l$ are situated p_l, respectively, q_l zero rows, where $\sum_{l=1}^{N} n_l = n$, $p_1 = q_N = 0$ and $p_l + k_l + q_l = k$ for all $l = 1, 2, \cdots, N$. Furthermore, assume that

$$(3) \qquad p_l < p_{l+1} < p_l + k_l, \qquad l = 1, 2, \cdots, N-1$$

and

$$(4) \qquad q_l < q_{l-1} < q_l + k_l, \qquad l = 2, 3, \cdots, N.$$

Considering the replacement of (3) by $p_l \leqq p_{l+1} \leqq p_l + k_l$ we observe that (a) elimination of $p_l = p_{l+1}$ by joining the submatrices involved decreases the complexity of decoding (see (30)) and (b) $p_{l+1} = p_l + k_l$ implies that a segment of length p_{l+1} of the message word and its remaining segment are encoded, and therefore may be decoded, separately. Relaxation of (4) is abandoned for similar reasons. It proves useful to define $p_{N+1} = p_N + k_N = k$ and $q_0 = k_1 + q_1 = k$. By this convention and (3)–(4) all the increments

$$(5) \qquad \Delta p_l = p_{l+1} - p_l, \quad \Delta q_l = q_{l-1} - q_l, \quad l = 1, 2, \cdots, N$$

are positive ($\Delta p_N = k_N$ and $\Delta q_1 = k_1$).

A recursive decoding algorithm based on the Hadamard transform for a code with generator matrix G given by (2) is developed in § 3. In subsequent sections, application to convolutional codes, high rate block codes and combined codes are discussed. A simple example is provided in § 7.

In all the cases considered the algorithm has a pronounced structural regularity, a feature particularly important for VLSI implementation. The Viterbi algorithm, when applied to block and combined codes [6], [7], [9], [10], is less regular in general. On the other hand, both the computational complexity and the memory requirement of the two algorithms turn out to be equal.

Application of fast Hadamard transform for maximum likelihood soft decision decoding was introduced in [4] (see also [8] and [5, p. 419]), where first order Reed–Muller codes are considered. In [2], a generalization to binary block and convolutional codes is described. A more efficient utilization of the Hadamard transform for decoding block codes with low to moderate dimensions, by exploiting the existence of certain kinds of codewords, is presented in [1]. This technique, applied to each $G^{(l)}$ in (2) may, in principle, result in substantial savings at the expense of regularity. However, no such reduction of complexity is apparently available for a typical code of the kind considered in §§ 4–6.

2. Preliminaries. For any $\mathbf{y} = (y_0, y_1, \cdots, y_{i-1}) \in \mathrm{GF}(2)^i$ let $b(\mathbf{y}) = \sum_{j=0}^{i-1} y_j 2^j$. The all-ones vector of length 2^i is denoted $\mathbf{E}_i$, and $\mathbf{E}$ is an all-ones vector of unspecified length. H_i is the $2^i \times 2^i$ naturally ordered ([5, p. 44]) Hadamard matrix, and $\otimes$ stands for the Kronecker product.

Consider first the following partitioning of G:

$$G = (G^1 G^2 \cdots\cdots\cdots G^N) \tag{6}$$

where G^l is a $k \times n_l$ submatrix; $l = 1, 2, \cdots, N$. Let $\{\mathbf{g}_i^l\}$ be the columns of G^l, and let

$$u_j^l = \begin{cases} \sum_{i \in A_j^l} \mu(v_i), & A_j^l \neq \varnothing, \\ 0, & A_j^l = \varnothing \end{cases} \tag{7}$$

where $A_j^l = \{i \colon b(\mathbf{g}_i^l) = j\}$ for $j = 0, 1, \cdots, 2^k - 1$. Denoting

$$M^l(\mathbf{s}) = \sum_{j=0}^{2^k-1} (-1)^{\langle \mathbf{s}, \mathbf{g}_j^{l*} \rangle} u_j^l$$

where $\mathbf{g}_j^{l*} \in \mathrm{GF}(2)^k$ is defined by $b(\mathbf{g}_j^{l*}) = j$, (1) becomes $M(\mathbf{s}) = \sum_{l=1}^N M^l(\mathbf{s})$. Let

$$\mathbf{U}^l = \mathbf{u}^l H_k, \qquad l = 1, 2, \cdots, N$$

where

$$\mathbf{u}^l = (u_0^l, u_1^l, \cdots, u_{2^k-1}^l).$$

The following result is immediate.

LEMMA 1. *For a code* $\mathscr{C}$ *with generator matrix given by* (6), $M(\mathbf{s})$ *given by* (1) *is equal to the jth component, where* $j = b(\mathbf{s})$, *of* $\mathbf{U} = \sum_{l=1}^N \mathbf{U}^l$.

Now assume that G is partitioned as in (2), *i.e.*,

$$G^l = \begin{bmatrix} 0 \\ G^{(l)} \\ 0 \end{bmatrix}, \qquad l = 1, 2, \cdots, N$$

where the upper and lower 0 *submatrices have* p_l *and* q_l *rows, respectively. Denote* $\mathbf{u}^{(l)} = (u_0^{(l)}, u_1^{(l)}, \cdots, u_{2^{k_l}-1}^{(l)})$, *where* $u_j^{(l)}$ *is defined similarly to* u_j^l *in* (7) *according to the columns* $\{\mathbf{g}_i^{(l)}\}$ *of* $G^{(l)}$, *and write* $\mathbf{U}^{(l)} = \mathbf{u}^{(l)} H_{k_l}$.

LEMMA 2. $\mathbf{U}^l$ *and* $\mathbf{U}^{(l)}$ *satisfy the following relationship*:

$$\mathbf{U}^l = \mathbf{E}_{q_l} \otimes \mathbf{U}^{(l)} \otimes \mathbf{E}_{p_l}.$$

Proof. Write e_t for the vector with entry 1 in its 0th (leftmost) location and zero in its remaining $2^t - 1$ locations. Obviously $b(\mathbf{g}_i^l) = b(\mathbf{g}_i^{(l)}) \cdot 2^{p_l} < 2^{k_l + p_l}$. Therefore, regardless of $G^{(l)}$, $A_j^l = \varnothing$ for all $j = 0, 1, \cdots, 2^k - 1$ such that $j \neq j' 2^{p_l}$ where $j' = 0, 1, \cdots, 2^{k_l} - 1$. Thus, it follows that $\mathbf{u}^l = \mathbf{e}_{q_l} \otimes \mathbf{u}^{(l)} \otimes \mathbf{e}_{p_l}$. Hence, by a known property of the Hadamard matrix,

$$\mathbf{U}^l = (\mathbf{e}_{q_l} \otimes \mathbf{u}^{(l)} \otimes \mathbf{e}_{p_l})(H_{q_l} \otimes H_{k_l} \otimes H_{p_l}).$$

Due to the associative law of the Kronecker product and the property

$$(A \otimes B)(C \otimes D) = (AD) \otimes (BD)$$

([5, p. 421]) it follows that

$$\mathbf{U} = (\mathbf{e}_{q_l} H_{q_l}) \otimes (\mathbf{u}^{(l)} H_{k_l}) \otimes (\mathbf{e}_{p_l} H_{p_l}),$$

thus completing the proof.

COROLLARY. $M(\mathbf{s})$ *given by* (1) *is equal to the jth component, where* $j = b(\mathbf{s})$, *of*

$$\mathbf{U} = \sum_{l=1}^{N} \mathbf{E}_{q_l} \otimes \mathbf{U}^{(l)} \otimes \mathbf{E}_{p_l}. \tag{8}$$

DEFINITION. Denote $K_t = \{0, 1, \cdots, t-1\}$. For a positive integer ν and a positive divisor r thereof, let the operator of maxima evaluation over segments of length r, Me_r: $R^\nu \to R^{\nu/r}$, be defined by

$$\mathrm{Me}_r(\mathbf{V}) = (V_{j_0}, V_{j_1}, \cdots, V_{j_{(\nu/r)-1}})$$

where

$$V_{j_l} = \max\{V_i : i = lr, lr+1, \cdots, (l+1)r - 1\}, \qquad l = 0, 1, \cdots, (\nu/r) - 1. \tag{9}$$

Also, let corresponding operators of locations designation Ld_r^*: $R^\nu \to K_\nu^{\nu/r}$ and Ld_r: $R^\nu \to K_r^{\nu/r}$ be given, respectively, by

$$\mathrm{Ld}_r^*(\mathbf{V}) = (j_0, j_1, \cdots, j_{(\nu/r)-1}) \quad \text{and} \quad \mathrm{Ld}_r(\mathbf{V}) = \mathrm{Ld}_r^*(\mathbf{V}) (\mathrm{mod}\ r),$$

where each j_l satisfies (9) and is otherwise arbitrarily selected (in case the maximum is attained at several locations).

For $\mathbf{V}$ of length ν, $\mathrm{Me}_\nu(\mathbf{V})$ is the value of a maximal component of $\mathbf{V}$, and both $\mathrm{Ld}_\nu^*(\mathbf{V})$ and $\mathrm{Ld}_\nu(\mathbf{V})$ represent the location of such a component. Also, $\mathrm{Me}_1(\mathbf{V}) = \mathbf{V}$, $\mathrm{Ld}_1(\mathbf{V}) = \mathbf{0}$ and $\mathrm{Ld}_1^*(\mathbf{V}) = (0, 1, \cdots, \nu - 1)$. The following properties are easily verified.

(A) For a divisor r of ν

$$\mathrm{Me}_r(\mathbf{E} \otimes \mathbf{V}) = \mathbf{E} \otimes \mathrm{Me}_r(\mathbf{V}) \tag{10}$$

and

$$\mathrm{Ld}_r(\mathbf{E} \otimes \mathbf{V}) = \mathbf{E} \otimes \mathrm{Ld}_r(\mathbf{V}). \tag{11}$$

(B) For positive integers r_1 and r_2 such that $r_1 r_2$ divides ν

$$\mathrm{Me}_{r_1 r_2}(\mathbf{V}) = \mathrm{Me}_{r_1}[\mathrm{Me}_{r_2}(\mathbf{V})], \tag{12}$$

and it is possible to resolve the arbitrariness of $\mathrm{Ld}_{r_1}[\mathrm{Me}_{r_2}(\mathbf{V})]$ and $\mathrm{Ld}_{r_2}(\mathbf{V})$ for a specified

$\mathrm{Ld}_{r_1 r_2}(\mathbf{V})$ such that

$$\mathrm{Ld}_{r_1 r_2}(\mathbf{V}) = \mathrm{Ld}_{r_1}[\mathrm{Me}_{r_2}(\mathbf{V})] \cdot r_2 + \{\mathrm{Ld}_{r_2}(\mathbf{V})\}_{\mathrm{Ld}_{r_1}[\mathrm{Me}_{r_2}(\mathbf{V})]} \tag{13}$$

where $\{\mathbf{X}\}_{\mathbf{Z}}$ for any $\mathbf{X} \in R^{\eta}$ and $\mathbf{Z} \in K_r^{\eta/r}$ is given by

$$\{\mathbf{X}\}_{\mathbf{Z}} = (X_{Z_0}, X_{r+Z_1}, X_{2r+Z_2}, \cdots, X_{\eta - r + Z_{(\eta/r)-1}}).$$

(C) For $\mathbf{X}$ and $\mathbf{Z}$ as above

$$\{\mathbf{E} \otimes \mathbf{X}\}_{\mathbf{Z}} = \{\mathbf{X}\}_{\mathbf{Z}(\mathrm{mod}\ \eta)} \tag{14}$$

and

$$\{\mathbf{E} \otimes \mathbf{X}\}_{\mathbf{E} \otimes \mathbf{Z}} = \mathbf{E} \otimes (\{\mathbf{X}\}_{\mathbf{Z}}). \tag{15}$$

Note that (13) is a useful expansion of $\mathrm{Ld}_{r_1 r_2}(\mathbf{V})$ whenever all possible resolutions are equivalent. Computation of $\mathrm{Me}_r(\mathbf{V})$ requires $(\nu/r)(r-1)$ additions. Since

$$\frac{\nu}{r_1 r_2}(r_1 r_2 - 1) = \frac{\nu}{r_2}(r_2 - 1) + \frac{\nu}{r_1 r_2}(r_1 - 1),$$

evaluation of the two sides of (12) bear the same computational complexity. By induction, maximization performed in several steps, at each step over segments of equal length that constitute the whole vector, does not alter the complexity.

3. A recursive algorithm. For convenience we shall write

$$\mathrm{Me}_{2^t} = M_t, \quad \mathrm{Ld}^*_{2^t} = L^*_t, \quad \mathrm{Ld}_{2^t} = L_t.$$

Expressed in the new notation, properties (10)–(14) retain their shape except for obvious changes, such as products of subscripts being replaced by their sum.

LEMMA 3. *Let* $\mathbf{V}^{(1)} \in R^{2^a}$ *and* $\mathbf{V}^{(2)} \in R^{2^b}$ *where* $a \leqq b$. *Then for* $p = b - a$ *and any integer* r *such that* $p \leqq r \leqq b$

$$M_r(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}) = M_{r-p}(\mathbf{V}^{(1)} + M_p(\mathbf{V}^{(2)})). \tag{16}$$

Furthermore, for a specified $L_r(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)})$ *the arbitrariness of* $L_{r-p}(\mathbf{V}^{(1)} + M_p(\mathbf{V}^{(2)}))$ *and* $L_p(\mathbf{V}^{(2)})$ *may be resolved such that*

$$L_r(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}) = L_{r-p}(\mathbf{V}^{(1)} + M_p(\mathbf{V}^{(2)})) \cdot 2^p + \{L_p(\mathbf{V}^{(2)})\}_{L_{r-p}(\mathbf{V}^{(1)} + M_p(\mathbf{V}^{(2)}))}. \tag{17}$$

Proof. According to (12)

$$M_r(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}) = M_{r-p}(M_p(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)})).$$

It is straightforward to check that

$$M_p(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}) = \mathbf{V}^{(1)} + M_p(\mathbf{V}^{(2)}), \tag{18}$$

yielding (16). By (13)

$$L_r(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}) = L_{r-p}(M_p(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)})) \cdot 2^p$$
$$+ \{L_p(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)})\}_{L_{r-p}(M_p(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}))},$$

and (17) follows by (18) and

$$L_p(\mathbf{V}^{(1)} \otimes \mathbf{E}_p + \mathbf{V}^{(2)}) = L_p(\mathbf{V}^{(2)}).$$

ALGORITHM.

(a) Evaluate $M_k(\mathbf{U}) = M_{k_N}(\mathbf{W}^{(N)})$ with the aid of the forward recursion

$$\begin{aligned} &\mathbf{W}^{(1)} = \mathbf{U}^{(1)}, \\ &\mathbf{W}^{(l)} = \mathbf{U}^{(l)} + \mathbf{E}_{\Delta q_l} \otimes M_{\Delta p_{l-1}}(\mathbf{W}^{(l-1)}), \qquad l = 2, 3, \cdots, N \end{aligned} \tag{19}$$

and store the location vectors $\{L_{\Delta p_l}(\mathbf{W}^{(l)})\}$ (or $\{L^*_{\Delta p_l}(\mathbf{W}^{(l)})\}$).

(b) Trace back for $L_k(\mathbf{U}) = I_1$ using the recursion

$$\begin{aligned} &I_N = L_{k_N}(\mathbf{W}^{(N)}), \\ &I_l = I_{l+1} \cdot 2^{\Delta p_l} + \{L_{\Delta p_l}(\mathbf{W}^{(l)})\}_{I_{l+1}(\text{mod } 2^{k_l - \Delta p_l})}, \qquad l = N-1, N-2, \cdots, 1 \end{aligned} \tag{20}$$

and set $\mathbf{s}$ to be the k-bit radix-2 expansion of I_1. Alternatively, proceed according to the following backward recursion:

$$J_N = L^*_{k_N}(\mathbf{W}^{(N)}), \tag{21a}$$

$$J_l = \{L^*_{\Delta p_l}(\mathbf{W}^{(l)})\}_{J_{l+1}(\text{mod } 2^{k_l - \Delta p_l})}, \qquad l = N-1, N-2, \cdots, 1, \tag{21b}$$

and write

$$\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N) \tag{22}$$

where $\mathbf{s}_l$ is the Δ_{p_l} bit radix-2 expansion of

$$i_l = J_l(\text{mod } 2^{\Delta p_l}), \qquad l = N, N-1, \cdots, 1. \tag{23}$$

Proof. Referring to (8), let

$$\mathbf{Y}^{(l)} = M_{p_{l+1}}\left(\sum_{j=1}^{l} \mathbf{E}_{q_j - q_l} \otimes \mathbf{U}^{(j)} \otimes \mathbf{E}_{p_j}\right), \qquad l = 1, 2, \cdots, N.$$

Obviously $\mathbf{Y}^{(N)} = M_k(\mathbf{U})$. By writing

$$\mathbf{Y}^{(l)} = M_{p_{l+1}}\left(\mathbf{U}^{(l)} \otimes \mathbf{E}_{p_l} + \sum_{j=1}^{l-1} \mathbf{E}_{q_j - q_l} \otimes \mathbf{U}^{(j)} \otimes \mathbf{E}_{p_j}\right),$$

applying (16), the relationship $q_j - q_l = (q_j - q_{l-1}) + \Delta q_l$ and (10), it follows that $\mathbf{Y}^{(l)} = M_{\Delta p_l}(\mathbf{U}^{(l)} + \mathbf{E}_{\Delta q_l} \otimes \mathbf{Y}^{(l-1)})$. This yields an N-step recursive procedure for evaluating $M_k(\mathbf{U})$:

$$\begin{aligned} &\mathbf{Y}^{(0)} = \mathbf{0}, \\ &\mathbf{Y}^{(l)} = M_{\Delta p_l}(\mathbf{U}^{(l)} + \mathbf{E}_{\Delta q_l} \otimes \mathbf{Y}^{(l-1)}), \qquad l = 1, 2, \cdots, N. \end{aligned} \tag{24}$$

Denote

$$\mathbf{W}^{(l)} = \mathbf{U}^{(l)} + \mathbf{E}_{\Delta q_l} \otimes \mathbf{Y}^{(l-1)}, \qquad l = 1, 2, \cdots, N. \tag{25}$$

Then $\mathbf{Y}^{(l)} = M_{\Delta p_l}(\mathbf{W}^{(l)})$, concluding the proof of step (a). Now let

$$I_l = L_{k - p_l}\left[\sum_{j=l}^{N} \mathbf{E}_{q_j} \otimes \mathbf{U}^{(j)} \otimes \mathbf{E}_{p_j - p_l} + \mathbf{E}_{q_{l-1}} \otimes \mathbf{Y}^{(l-1)}\right], \qquad l = 1, 2, \cdots, N. \tag{26}$$

Then $I_1 = L_k(\mathbf{U})$, the location of a maximal component of $\mathbf{U}$. In view of the corollary, $\mathbf{s}$ is thus obtainable by expanding I_1 in radix-2. Proof of the recursion (20) is deferred to the Appendix. Consider the partitioning (22) of $\mathbf{s}$, where $\mathbf{s}_l$ has length Δp_l, and let $i_l = b(\mathbf{s}_l)$. Since each component of $L_{\Delta p_l}(\mathbf{W}^{(l)})$ is smaller than $2^{\Delta p_l}$, it follows by (20) that

$i_N = I_N$,

$$i_l = \{L_{\Delta p_l}(\mathbf{W}^{(l)})\}_{I_{l+1} (\mathrm{mod}\, 2^{k_l - \Delta p_l})}, \qquad l = 1, 2, \cdots, N-1 \tag{27}$$

and

$$I_l = I_{l+1} \cdot 2^{\Delta p_l} + i_l, \qquad l = 1, 2, \cdots, N-1. \tag{28}$$

Define

$$J_l = \{L^*_{\Delta p_l}(\mathbf{W}^{(l)})\}_{I_{l+1} (\mathrm{mod}\, 2^{k_l - \Delta p_l})}, \qquad l = N-1, N-2, \cdots, 1 \tag{29}$$

and J_N is given by (21a). Then (23) obviously holds; in particular $i_N = J_N = I_N$. By (27) and (28)

$$I_l = \lfloor I_{l+1} \cdot 2^{-(k_l - \Delta p_l)} \rfloor 2^{k_l - \Delta p_l} \cdot 2^{\Delta p_l} + J_l$$

and, due to (4), $k_{l-1} - \Delta p_{l-1} < k_l$. Consequently (21b) and (29) are the same, thus concluding the proof.

Application of this algorithm is straightforward and simple once the submatrices are determined in a generator matrix of $\mathscr{C}$ with double-echelon structure (2). This is demonstrated in § 7. Notice that the trellis-representation of a code, used for application of the Viterbi algorithm, is also nonunique [6], [10].

The computational complexity A of the algorithm is dominated by (19); hence it is very closely

$$A = k_1 2^{k_1} + \sum_{l=2}^{N} \left[k_l 2^{k_l} + \frac{2^{k_{l-1}}}{2^{\Delta p_{l-1}}} (2^{\Delta p_{l-1}} - 1) \right] + (2^{k_N} - 1) \tag{30}$$

addition-equivalent operations. Since $\frac{1}{2} \leqq (2^{\Delta p_l} - 1)/2^{\Delta p_l} < 1$ for all $l = 1, 2, \cdots, N-1$ due to (3), the following tight bound is obtained:

$$A < \sum_{l=1}^{N} (k_l + 1) 2^{k_l}. \tag{31}$$

Frequently $n_l < k_l$. Then direct, rather than fast, transform evaluation of $\mathbf{U}^{(l)}$ is preferable, allowing replacement of $k_l 2^{k_l}$ in (30) by $(n_l - 1) 2^{k_l}$ and even by $(n_l - 1) 2^{n_l}$ (the latter since entries in $\mathbf{U}^{(l)}$ are repeated). Thus for the case $n_l < k_l$, $l = 1, 2, \cdots, N$ we conclude that

$$A < \sum_{l=1}^{N} n_l 2^{k_l}, \tag{32}$$

and

$$A \approx \sum_{l=1}^{N} 2^{k_l}. \tag{33}$$

Furthermore, in this case the memory requirement, dominated by the length of the longest vector in (19) and by the lengths of the location vectors in either (20) or (21), is nearly

$$B = 2^{(k - \Delta p)_{\max}} + \sum_{l=1}^{N-1} 2^{k_l - \Delta p_l} \tag{34}$$

symbols, where $(k - \Delta p)_{\max} = \max \{k - \Delta p_l : l = 1, 2, \cdots, N-1\}$. Evidently,

$$B < N \cdot 2^{(k - \Delta p)_{\max}} \leqq N \cdot 2^{k_{\max} - 1}$$

where $k_{\max} = \max \{k_l\colon l = 1, 2, \cdots, N\}$. When fast Hadamard transform is used, then the first term in the right side of (34) has to be replaced by $2^{k_{\max}}$, and $B < N \cdot 2^{k_{\max}}$. In summary, compared to the basic algorithm in [1], a computational complexity gain as well as memory reduction of $O(2^{k-k_{\max}})$ is achieved.

It seems difficult to obtain a recursion expressed in terms of $\{i_l\}$. However assume, for example, that $k_l < \Delta p_l + \Delta p_{l+1} + \Delta p_{l+2}$ for all $l = 1, 2, \cdots, N-2$. Then repeated use of (28) and substitution into (27) yield

$$\begin{aligned} &i_N = L_{k_N}(\mathbf{W}^{(l)}), \\ (35) \qquad &i_{N-1} = \{L_{\Delta p_{N-1}}(\mathbf{W}^{(N-1)}\}_{i_N (\operatorname{mod} 2^{k_{N-1}-\Delta p_{N-1}})}, \\ &i_l = \{L_{\Delta p_l}(\mathbf{W}^{(l)})\}_{j_l}, \qquad l = N-2, N-3, \cdots, 1 \end{aligned}$$

where

$$j_l = i_{l+2} 2^{\Delta p_{l+1}} + i_{l+1} (\operatorname{mod} 2^{k_l - \Delta p_l}).$$

4. Decoding of convolutional codes. Let $\mathscr{C}$ be the rth truncation of a binary (η, κ) convolutional code with memory m and polynomial generator matrix $\sum_{i=0}^{m} G_i x^i$. Then a κr by $\eta(r+m)$ binary generator matrix of $\mathscr{C}$ is given by

$$(36) \qquad G = \left[\begin{array}{ccccccc} G_0 & G_1 \cdots\cdots G_m & & & & & \\ & G_0 & G_1 \cdots\cdots G_m & & & 0 & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & 0 & & G_0 & G_1 \cdots\cdots G_m & & \\ & & & & G_0 & G_1 \cdots\cdots G_m & \end{array}\right]$$

and, assuming $r > m + 1$, it obviously admits the representation (2) with the following parameters: $N = r - m$; $k_l = \kappa(m+1)$ for $l = 1, 2, \cdots, N$; $n_l = \eta$ for $l = 2, 3, \cdots, N-1$ and $n_1 = n_L = \eta(m+1)$; $\Delta p_l = \kappa$ for $l = 1, 2, \cdots, N-1$ and $\Delta q_l = \kappa$ for $l = 2, 3, \cdots, N$. In particular, $(G^{(l)})^T = (G_m^T G_{m-1}^T \cdots G_0^T)$ for all $l = 2, 3, \cdots, N-1$, where a superscript T indicates transposition. A time-varying convolutional code is representable by a matrix structured similarly to (36), and consequently by the same values of parameters, but $G^{(l)}$ varies with l in general. Version (35) of step (b) is applicable if $m \leqq 2$; otherwise a suitable generalization of it may be adopted.

According to (31)

$$A < (r-m)(\kappa m + \kappa + 1) 2^{\kappa(m-1)}.$$

Typically for good convolutional codes, the memory m exceeds the inverse of the rate κ/η. In this case, $n_l < k_l$ for all $l = 2, 3, \cdots, N-1$ and the following counterparts of (32) and (33), respectively, are obtained:

$$(37) \qquad A < \eta(r+m) 2^{\kappa(m+1)},$$

$$(38) \qquad A \approx (r-m+2) 2^{\kappa(m+1)}$$

where for obtaining (38) the zero-zones of $G^{(1)}$ and $G^{(N)}$ must also be exploited. The complexity of the straightforwardly applied Viterbi algorithm is given approximately by the right-hand side of (37), but a closer look [3] at the trellis diagram reveals possible simplifications that lead to (38). The memory requirements of the Viterbi algorithm and the algorithm derived here are also approximately equal.

5. Decoding of high rate block codes. Let $\mathscr{C}$ be a (n, k) binary linear block code such that $k < n < 2k - 1$. It can be shown that there exists (a generally nonunique) equivalent code generated by

$$(39)\qquad G=\begin{bmatrix}
1x\cdots xx & & & & \\
& 1x\cdots xx & & & \\
& \cdot\cdot\quad\cdot\cdot & & & \\
& & \cdot\cdot\quad\cdot\cdot & & 0\\
& & \cdot\cdot\quad\cdot\cdot & & \\
& & & \cdot\cdot\quad\cdot\cdot & \\
& & & \cdot\cdot\quad\cdot\cdot & \\
0 & & & \cdot\cdot\quad\cdot\cdot & \\
& & & & \cdot\cdot\quad\cdot\cdot \\
& & & & 1x\cdots xx \\
& & & & \quad 1x\cdots x1
\end{bmatrix}$$

where each x stands for either 0 or 1, irrespective of the value of other entries denoted by x. Such a matrix corresponds to a time-varying version of (36) with 1 substituted for both η and κ, k for r and $n - k$ for m. A time-invariant version results whenever $\mathscr{C}$ is cyclic.

Thus, aside from post-multiplication by a constant matrix (and pre-permutation of labels), the complexity is given by

$$(40)\qquad A \approx (2k-n+2)^{n-k+1}.$$

Since $n_l = 1$, except for $l = 1$ and $l = 2k - n$, (40) is practically an equality. The algorithm derived in [10] has about the same complexity and memory requirement as the algorithm presented here, but the latter is distinguished by a structural regularity not found in [10]. Regularity is particularly important for VLSI implementation. For high rate but rather short codes the approach of [1] may prove to be more efficient.

6. Decoding of combined codes. Consider first a product code $\mathscr{C}$ with generator matrix

$$(41)\qquad G = G_a \otimes G_b$$

where G_a and G_b are, respectively, the generator matrices of a binary (n_a, k_a) code $\mathscr{C}_a$ and a binary (n_b, k_b) code $\mathscr{C}_b$. It is possible to use Algorithm A of [1] if $2k_ak_b \leqq n_an_b + 1$, with complexity $k_ak_b2^{k_ak_b}$, whereas, if $2k_ak_b > n_an_b + 1$, we may resort to the algorithm derived here, with complexity $(2k_ak_b - n_an_b + 2)2^{n_an_b - k_ak_b + 1}$. However, assuming that $k_a < n_a < 2k_a - 1$, select for $\mathscr{C}_a$ a generator matrix that possesses the pattern of (39). Then G given by (41) has the double-echelon structure (2) with the following parameters: $N = 2k_a - n_a$; $n_1 = n_N = n_b(n_a - k_a + 1)$ and $n_l = n_b$ for $l = 2, 3, \cdots, N - 1$; $k_l = k_b(n_a - k_a + 1)$ for $l = 1, 2, \cdots, N$; $\Delta p_l = k_b$ for $l = 1, 2, \cdots, N - 1$ and $\Delta q_l = k_b$ for $l = 2, 3, \cdots, N$. Hence by (31)

$$A < (2k_a - n_a)[k_b(n_a - k_a + 1) + 1]2^{k_b(n_a - k_a + 1)}.$$

For the frequently encountered case where $n_b < k_b(n_a - k_a + 1)$ we obtain $A \approx (2K_a - n_a + 2)2^{k_b(n_a - k_a + 1)}$. If $\mathscr{C}_b$, rather than $\mathscr{C}_a$, is a high rate code, then write $G_a \otimes G_b = P_b(G_b \otimes G_a)P_a$ where P_a and P_b are permutation matrices. Reduced complexity thus results by adopting $G_b \otimes G_a$, a columns and rows permuted version $P_b^{-1}GP_a^{-1}$ of G, as generator matrix.

Now let $\mathscr{C}_a$ be an (n_a, k_a) binary code with generator matrix G'_a, and let $\mathscr{C}_b$ be an (n_b, k_b) q-ary code, where $q = 2^{k_a}$, with a *binary* $k_a k_b$ by $k_a n_b$ generator matrix G_b. Denote $G_a = I \otimes G'_a$ where I is an n_b-dimensional unity matrix. A binary $(n_a n_b, k_a k_b)$ concatenated code $\mathscr{C}$ is generated by

$$G = G_b G_a. \tag{42}$$

The practically interesting case is: the inner code $\mathscr{C}_a$ is of low rate whereas the outer code $\mathscr{C}_b$ has high rate, i.e., $2k_a \leqq n_a + 1$ and $2k_b > n_b + 1$. Then we may assume that G_b has the pattern of (39). Hence, G given by (42) possesses the double-echelon structure (2) with the following parameters: $N = 2k_b - n_b$; $n_1 = n_N = n_a(n_b - k_b + 1)$ and $n_l = n_a$ for $l = 2, 3, \cdots, N-1$; $k_l = k_a(n_b - k_b + 1)$ for $l = 1, 2, \cdots, N$; $\Delta p_l = k_a$ for $l = 1, 2, \cdots, N-1$ and $\Delta q_l = k_a$ for $l = 2, 3, \cdots, N$. The computational complexity of decoding, assuming $n_a < k_a(n_b - k_b + 1)$, is approximately $(2k_b - n_b + 2)q^{n_b - k_b + 1}$.

Comparison of the results presented in this section with the corresponding algorithm in [7] and [10] reveals nearly the same complexity and memory requirement.

7. An example. Consider a code $\mathscr{C}$ generated by

$$G = \begin{pmatrix} 1101 & & & \\ 1101 & & & 0 \\ 1101 & & & \\ 0 & 111 & & \\ & & 1111 & \end{pmatrix}. \tag{43}$$

There are several ways of partitioning this matrix in conformity with the double-echelon structure defined by (2). Let us adopt the following:

$$G^{(1)} = \begin{pmatrix} 11010 \\ 01101 \\ 00110 \end{pmatrix}, \tag{44a}$$

$$G^{(2)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{44b}$$

$$G^{(3)} = \begin{pmatrix} 1100 \\ 1111 \end{pmatrix}. \tag{44c}$$

Then $k_1 = 3$, $k_2 = k_3 = 2$, $\Delta p_1 = 2$, $\Delta p_2 = 1$, $\Delta p_3 = 2$, $\Delta q_1 = 3$ and $\Delta q_2 = \Delta q_3 = 1$. Assume that application of the log-likelihood ratio $\mu(\cdot)$ to each entry of the received vector $\mathbf{v}$ yields the following values:

$$(3, -4, 6, 2, -3|5|9, -7, -1, 8), \tag{45}$$

listed in an order identical to the order of the entries of $\mathbf{v}$. The partitioning in (45), indicated by vertical lines, corresponds to the partitioning of G expressed by (44a)–(44c). Now, by the definition that precedes Lemma 2,

$$\mathbf{u}^{(1)} = (0, 3, -3, -4, 0, 2, 6, 0),$$

$$\mathbf{u}^{(2)} = (0, 0, 0, 5),$$

$$\mathbf{u}^{(3)} = (0, 0, 7, 2).$$

Hence

$$\mathbf{U}^{(1)} = \mathbf{u}^{(1)}H_3 = (4,2,6,-12,-12,-6,14,4),$$
$$\mathbf{U}^{(2)} = \mathbf{u}^{(2)}H_2 = (5,-5,-5,5),$$
$$\mathbf{U}^{(3)} = \mathbf{u}^{(3)}H_2 = (9,5,-9,-5).$$

According to (19)

$$\mathbf{W}^{(1)} = (4,2,6,-12,|-12,-6,14,4),$$
$$\mathbf{W}^{(2)} = (5,-5,-5,5)+(1,1)\otimes(6,14) = (11,9|1,19),$$
$$\mathbf{W}^{(3)} = (9,5,-9,-5)+(1,1)\otimes(11,19) = (20,24,2,14),$$

where vertical lines mark the separation into the segments over which the maxima were evaluated. Since $L_2(\mathbf{W}^{(1)}) = (2,2)$, $L_1(\mathbf{W}^{(2)}) = (0,1)$ and $L_2(\mathbf{W}^{(3)}) = 1$, it follows by (20) that

$$I_3 = 1,$$
$$I_2 = 1\cdot 2 + \{(0,1)\}_1 = 3,$$
$$I_1 = 3\cdot 2^2 + \{(2,2)\}_{3 (\mathrm{mod}\ 2)} = 14.$$

Consequently,

$$\mathbf{s} = (0,1,1,1,0). \tag{46}$$

Noting that $L_2^*(\mathbf{W}^{(1)}) = (2,6)$, $L_1^*(\mathbf{W}^{(2)}) = (0,3)$ and $L_2^*(\mathbf{W}^{(3)}) = 1$, the alternative inverse recursion (21) yields

$$J_3 = 1,$$
$$J_2 = \{(0,3)\}_1 = 3,$$
$$J_1 = \{(2,6)\}_{3 (\mathrm{mod}\ 2)} = 6.$$

Hence by (23) and (22)

$$\mathbf{s} = (01|1|10),$$

in agreement with (46).

We remark that, for the foregoing illustration, we deliberately selected a small code with its generator matrix partitioned such that the decoding procedure exhibits a slight structural irregularity. Notice, however, that this code is not of the types considered in §§ 4–6.

Appendix.

Proof of (20). By (26) $I_N = L_{k_N}(\mathbf{W}^{(N)})$. Due to $p_j - p_l = (p_j - p_{l+1}) + \Delta p_l$ and $q_{l-1} = q_l + \Delta q_l$

$$I_l = L_{k-p_l}\left(\left(\sum_{j=l+1}^{N} \mathbf{E}_{q_j} \otimes \mathbf{U}^{(j)} \otimes \mathbf{E}_{p_j - p_{l+1}} \right) \otimes \mathbf{E}_{\Delta p_l} + \mathbf{E}_{q_l} \otimes (\mathbf{U}^{(l)} + \mathbf{E}_{\Delta q_l} \otimes \mathbf{Y}^{(l-1)}) \right).$$

Thus by (14), (17) and (24)

$$I_l = L_{k-p_{l+1}}\left(\sum_{j=l+1}^{N} \mathbf{E}_{q_j} \otimes \mathbf{U}^{(j)} \otimes \mathbf{E}_{p_j - p_{l+1}} + \mathbf{E}_{q_l} \otimes \mathbf{Y}^{(l)} \right) \cdot 2^{\Delta p_l} + X_l$$
$$= I_{l+1} \cdot 2^{\Delta p_l} + X_l$$

where in view of (15) and (25)

$$X_l = \{\mathbf{E}_{q_l} \otimes L_{\Delta p_l}(\mathbf{W}^{(l)})\}_{I_{l+1}}.$$

The result now follows by (14).

Acknowledgments. The authors thank Moshe Dubiner and Aner Shalev for a number of helpful discussions.

REFERENCES

[1] Y. Be'ery and J. Snyders, *Optimal soft decision block decoders based on the fast Hadamard transform*, IEEE Trans. Inform. Theory, IT-32 (1986), pp. 355–364.

[2] G. C. Clark, Jr. and R. C. Davis, *A decoding algorithm for group codes and convolution codes based on the fast Fourier-Hadamard transform*, IEEE Internat. Symposium on Information Theory, Ellenville, NY, 1969.

[3] G. D. Forney, *The Viterbi algorithm*, Proc. IEEE, 61 (1973), pp. 268–278.

[4] R. R. Green, *A serial orthogonal decoder*, JPL Space Program Summary, No. 37-39, Vol. IV, pp. 247–253, 1966.

[5] F. K. MacWillams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1977.

[6] J. L. Massey, *Foundation and methods of channel coding*, Internat. Conference on Information Theory and Systems, NTG-Fachberichte, vol. 65, Berlin, 1978.

[7] J. F. Pieper, J. G. Proakis, R. R. Reed and J. K. Wolf, *Design of efficient coding and modulation for Rayleigh fading channel*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 457–468.

[8] E. C. Posner, *Combinatorial structures in planetary reconnaissance*, in Error Correcting Codes, H. B. Mann, ed., John Wiley, New York, 1969, pp. 15–46.

[9] G. Solomon and H. C. A. van Tilborg, *A connection between block and convolutional codes*, SIAM J. Appl. Math., 2 (1979), pp. 358–369.

[10] J. K. Wolf, *Efficient maximum likelihood decoding of linear block codes*, IEEE Trans. Inform. Theory. IT-24 (1978) pp. 76–80.

SIAM J. ALG. DISC. METH.
Vol. 8, No. 4, October 1987

ON ONE-SIDED JACOBI METHODS FOR PARALLEL COMPUTATION*

P. J. EBERLEIN†

Abstract. Convergence proofs are given for one-sided Jacobi/Hestenes methods for the singular value problem. The limiting form of the matrix iterates for the Hestenes method with optimization when the original matrix is normal is derived; this limiting matrix is block diagonal, where the blocks are multiples of unitary matrices. A variation in the algorithm to guarantee convergence to a diagonal matrix for the symmetric eigenvalue problem is shown. Implementation techniques for parallel computation, in particular, on the hypercube are indicated.

Key words. parallel computation, one-sided Jacobi methods, Hestenes method, multiprocessors, hypercube, singular values, eigenvalue problem

AMS(MOS) subject classifications. 65F10, 65H20, 65F05, 15

1. Introduction. We recently became interested in one-sided Jacobi methods for the parallel implementation of a program for finding the eigenvalues (and/or singular values) of an $n \times n$ (or $r \times n$) matrix on a hypercube. We had first considered using a block, or "patch," distribution of the data among the nodes. However, in order to compute both the left and right rotations of a similarity transformation, we were faced with the necessity of sending, for each rotation, small amounts of information both horizontally and vertically across an array of processors.

After writing programs for both one- and two-sided procedures, we decided to investigate further the one-sided algorithm developed originally by Hestenes [6], and cited most recently by Brent and Luk [1] for which a "strip" distribution of data is possible. The use of this strip data distribution of columns allows each processor to contain all the information needed to determine and to perform a rotation. This computation is, in addition, highly vectorizable. The columns of the matrix and the associated vectors may then be sent to other processors to perform subsequent calculations; a substantial amount of computation is thus performed in each processor before communication is necessary.

Hestenes was primarily interested in matrix inversion, but his biorthogonalization technique was naturally applicable to the singular value problem, not the symmetric eigenvalue problem. The Hestenes procedure, given a matrix A, finds an orthogonal matrix V such that AV has orthogonal columns. Hence, $(AV)^t(AV) = V^t(A^tA)V$ is diagonal. We proceeded with an implementation to obtain eigenvalues, hoping that V^tAV would be diagonal for those matrices, A, which we believed likely to occur in the applications in which we were interested. We soon found a set of matrices, from the chemical problem in which we were interested, for which this is not true. We investigate here the form that V^tAV does take in general.

As a result we find a new bound and proof of convergence, which includes an ordering of the singular values in the limiting matrix, as well as the form of the limiting matrix of V^tAV when the original matrix is square and normal.

Finally, we present a slight change in the rotational parameter which, in the symmetric case, ensures convergence to the diagonal matrix of eigenvalues, and which is amenable to implementation on multiprocessors, and on the hypercube in particular.

* Received by the editors August 4, 1986; accepted for publication (in revised form) July 30, 1987.

† Department of Computer Science, University at Buffalo, State University of New York, Buffalo, New York 14260.

Thus we are able to compute the actual Jacobi parameters at no greater cost than that of the Hestenes procedure. Finally, this technique is available to any Jacobi procedure which uses only local matrix information for the computation of the rotational angle, as, for example, in the nonsymmetric case.

A number of papers, besides that of Hestenes, have referred to, or reinvented, this one-sided Jacobi procedure over the years: Chartres [2], Eberlein [3], Kaiser [7], Nash [8], and most recently, Brent and Luk [1]. The motivation of the above-mentioned authors is varied. Brent and Luk cite the "Hestenes" procedure as a potential one for the determination of singular values in the parallel setting of a linear array of processors. Chartres, Eberlein and Nash recommended the procedure when external memory was a necessity, i.e., when the problem size exceeded the internal memory of the machines available. Kaiser appears to have believed that he had invented a new Jacobi-type procedure, the "JK Method" (Jacobi–Kaiser?) for finding the eigenvalues of a square symmetric matrix. Both Nash and Kaiser seem to have assumed that V^tAV would be diagonal when A is symmetric. Ordering of the roots is mentioned by Eberlein, Kaiser and Nash as a byproduct of regarding the procedure as an optimization problem.

The lack of interest in these procedures over the years is most certainly due to their poor performance on serial machines, for which they are not recommended.

2. The algorithm reviewed. We review the "one-sided" Jacobi algorithm, and establish notation. We consider the case of an $r \times n$ matrix A for which the singular values, or eigenvalues ($r = n$), and corresponding vectors are to be determined. We consider the sequence

$$A_{i+1} = A_i R_i,$$

where $A_0 = A$ and R_i is an n-dimensional plane rotation. We define an n-dimensional plane rotation, $R = (r_{i,j})$, by

$$\begin{aligned} & r_{k,k} = r_{m,m} = \cos\phi = c, \\ (1) \qquad & r_{m,k} = -r_{k,m} = \sin\phi = s, \qquad 1 \leqq k < m \leqq n, \\ & r_{i,j} = \delta_{i,j} \quad \text{for } i, j \neq k, m, \text{ where } \delta_{i,j} \text{ is the Kronecker delta.} \end{aligned}$$

Because all operations we use involve the columns of A, we represent the matrix A in terms of its columns: $A = (\mathbf{a}^1, \mathbf{a}^2, \cdots, \mathbf{a}^n)$. We consider a single transformation $A' = AR$. The multiplication of A on the right by a rotation R affects only the columns of the product A'. We have

$$(2) \qquad \begin{aligned} \mathbf{a}^{k'} &= c\,\mathbf{a}^k + s\,\mathbf{a}^m, & \mathbf{v}^{k'} &= c\mathbf{v}^k + s\mathbf{v}^m, \\ \mathbf{a}^{m'} &= -s\mathbf{a}^k + c\mathbf{a}^m, & \mathbf{v}^{m'} &= -s\mathbf{v}^k + c\mathbf{v}^m. \end{aligned}$$

Initially, we set $V = (\mathbf{v}^1, \mathbf{v}^2, \cdots, \mathbf{v}^n) = I$, and thereafter, $V' = VR$, as above. The matrix V is used to accumulate the product of rotations. (Note that the vectors $\mathbf{a}^i$ and $\mathbf{v}^i$ can be updated as a single vector.) The Hestenes angle for the rotation R is chosen to annihilate $(A^tA)_{k,m}$ as for the usual Jacobi method acting on the matrix A^tA. We have

$$(3) \qquad \tan 2\phi = 2\langle \mathbf{a}^k, \mathbf{a}^m \rangle / (\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2)$$

where $\langle \mathbf{a}^k, \mathbf{a}^m \rangle$ denotes the inner product between the vectors $\mathbf{a}^k$ and $\mathbf{a}^m$, and $\|\mathbf{a}^k\|^2 = \langle \mathbf{a}^k, \mathbf{a}^k \rangle$. As is well known, the equation for $\tan 2\phi$ has two solutions for $\tan \phi$. Letting

$\tan 2\phi = 1/\alpha$, then the smaller (or the larger) of the two solutions may be found by:

$$\begin{aligned} &t_s := 1/(|\alpha| + \sqrt{(1+\alpha^2)});\\ &\text{if } \alpha < 0, \text{ then } t_s := -t_s;\\ &\text{if (Large angle desired) then } t_L := -1/t_s;\\ &c := 1/\sqrt{(1+t^2)} \text{ and } s := t^*c. \end{aligned} \tag{4}$$

Note that the same angle may be regarded as maximizing $\|\mathbf{a}^{k'}\|^2$, provided that the smaller or larger angle is appropriately chosen. The potential use of the larger angle has been mentioned by Rutishauser [9], for sorting the eigenvalues, as well as by Eberlein [3] and Nash [8].

3. Convergence proofs. We introduce a lemma, which gives us a new bound for convergence in the theorem which follows.

LEMMA. *Let A be an $r \times n$ real matrix, $r \geqq n$, and R an n-dimensional plane rotation in the (k, m), $1 \leqq k < m \leqq n$ plane. Let $A' = AR$, and let* $\tan \phi$ *for the rotation be chosen so that $\langle \mathbf{a}^{k'}, \mathbf{a}^{m'} \rangle = 0$, and such that $\|\mathbf{a}^{k'}\|^2$ is maximized, i.e., if $\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2 < 0$, then t_L is used. Then*

$$(\|\mathbf{a}^{k'}\|^2 - \|\mathbf{a}^k\|^2) \geqq \langle \mathbf{a}^k, \mathbf{a}^m \rangle^2 / \max(\|\mathbf{a}^k\|^2, \|\mathbf{a}^m\|^2).$$

Proof. Because $\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2$ has the same sign as $\cos 2\phi$, straightforward computation leads to

$$\|\mathbf{a}^{k'}\|^2 = \tfrac{1}{2}[\|\mathbf{a}^k\|^2 + \|\mathbf{a}^m\|^2 + (\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2)\cos 2\phi + 2\sin 2\phi \langle \mathbf{a}^k, \mathbf{a}^m \rangle].$$

Using our choice of ϕ, we have

$$2(\|\mathbf{a}^{k'}\|^2 - \|\mathbf{a}^k\|^2) = \|\mathbf{a}^k\|^2 + \|\mathbf{a}^m\|^2 + \sqrt{[(\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2)^2 + 4\langle \mathbf{a}^k, \mathbf{a}^m \rangle^2]} - 2\|\mathbf{a}^k\|^2.$$

Letting $Q = \langle \mathbf{a}^k, \mathbf{a}^m \rangle$ and $P^2 = |\,\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2|$, we have

$$\begin{aligned} 2(\|\mathbf{a}^{k'}\|^2 - \|\mathbf{a}^k\|^2) &= \|\mathbf{a}^k\|^2 + \|\mathbf{a}^m\|^2 + \sqrt{(P^4 + 4Q^2)} - 2\|\mathbf{a}^k\|^2\\ &\geqq -P^2 + \sqrt{(P^4 + 4Q^2)} = 4Q^2/(P^2 + \sqrt{(P^4 + 4Q^2)}), \end{aligned}$$

because $(-P^2 + \sqrt{(P^4 + 4Q^2)})(P^2 + \sqrt{(P^4 + 4Q^2)}) = 4Q^2$. Hence,

$$2(\|\mathbf{a}^{k'}\|^2 - \|\mathbf{a}^k\|^2) \geqq 4Q^2/(|\,\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2| + \sqrt{(\|\mathbf{a}^k\|^4 - 2\|\mathbf{a}^k\|^2\|\mathbf{a}^m\|^2 + \|\mathbf{a}^m\|^4 + 4\langle \mathbf{a}^k, \mathbf{a}^m \rangle^2)}),$$

and, using the Schwartz inequality, $\langle \mathbf{a}^k, \mathbf{a}^m \rangle \leqq \|\mathbf{a}^k\|\,\|\mathbf{a}^m\|$, we obtain

$$\begin{aligned} (\|\mathbf{a}^{k'}\|^2 - \|\mathbf{a}^k\|^2) &\geqq 2\langle \mathbf{a}^k, \mathbf{a}^m \rangle^2/(|\,\|\mathbf{a}^k\|^2 - \|\mathbf{a}^m\|^2| + \|\mathbf{a}^k\|^2 + \|\mathbf{a}^m\|^2)\\ &= \langle \mathbf{a}^k, \mathbf{a}^m \rangle^2/\max(\|\mathbf{a}^k\|^2, \|\mathbf{a}^m\|^2). \end{aligned}$$

We use this lemma to prove the following theorem.

THEOREM 1. *Let A be a real $r \times n$ matrix, and let $A_i = AR_1R_2 \cdots R_i$. If each rotation is chosen to maximize the 2-norm of the kth column in A_i as described above, and the pivot pairs, (k, m), $1 \leqq k < m \leqq n$, are chosen so that every possible pair occurs in some regular order in every sweep, then the matrix AV, where $V = R_1R_2 \cdots R_i$, converges to a matrix, say B, such that $B^tB = V^tA^tAV = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$, the singular values of A, where $\sigma_1^2 \geqq \sigma_2^2 \geqq \cdots \geqq \sigma_n^2$.*

Proof. Let $S_j = \|\mathbf{a}^1\|^2 + \|\mathbf{a}^2\|^2 + \cdots + \|\mathbf{a}^j\|^2$ for j, $1 \leqq j \leqq n - 1$ denote the sum of squares of column norms for iterate A_i. We omit notation for i, though it should be understood. We will use the above lemma to show that each S_j forms a bounded monotone increasing sequence, which approaches a limit in which $\langle \mathbf{a}^j, \mathbf{a}^m \rangle = 0$, for all $j \neq m$, $1 \leqq j < m \leqq n$.

We note first that any rotation R with pivot pair (k, m), $k < m$, changes the norms of only the kth and mth columns. The kth column is increased as indicated in the lemma while the mth column is decreased by the same amount. Since rotations are orthogonal, the norm of the matrix AR remains unchanged and provides an upper bound for the S_j.

If $j = 1$, then S_1 is increased by any rotation with $(k, m) = (1, m)$ and remains unchanged by any other. S_2 is unchanged by the rotation R with $(k, m) = (1, 2)$, increased by a rotation with $(k, m) = (2, m)$, and left unchanged by the remainder. Similarly S_j is increased by rotations with pivots (j, m) where $m > j$; all other rotations leave S_j unchanged. Hence, if all possible pivot pairs appear in each sweep, all the S_j are monotone increasing and approach limits, in which, by the lemma, $|\langle \mathbf{a}^j, \mathbf{a}^m \rangle|$ approach zero $j < m \leqq n$. Letting $AV = B$, we have B^tB is diagonal. All the $\|\mathbf{b}_j\|^2$ also must approach limits since $S_1 = \|\mathbf{b}_1\|^2$, $S_2 = \|\mathbf{b}_1\|^2 + \|\mathbf{b}_2\|^2$ approaches a limit, and hence so does $\|\mathbf{b}_2\|^2$. This is similar for the others. We also have $\|\mathbf{b}_1\|^2 \geqq \|\mathbf{b}_2\|^2 \geqq \cdots \geqq \|\mathbf{b}_n\|^2$ because we have chosen the larger angle whenever $\|\mathbf{a}^j\|^2 < \|\mathbf{a}^m\|^2$, for $j < m$. The theorem follows.

We note that in the above argument no special ordering of pivots is used.

We now assume that A is square and normal and ask if $V^tB = V^tAV$ assumes any special form. In particular, we are interested in $V^tB = V^tAV$ when A is symmetric.

THEOREM 2. *Let the hypotheses of Theorem* 1 *hold but let A be a square normal matrix. Then $V^tB = V^tAV$ is block diagonal, where the blocks are multiples of orthogonal matrices.*

Proof. Let the matrix $B = AV$. Because the columns of B are orthogonal we have $B^tB = \operatorname{diag}(d_{i,i}^2)$, where $d_{1,1}^2 \geqq d_{2,2}^2 \geqq \cdots \geqq d_{n,n}^2$. Assume for the moment that $d_{n,n}^2 > 0$, and define $P = BD^{-1}$. P is orthogonal since $P^tP = I$. Now $B = PD$, and $A = PDV^t$. Since $A^t = VDP^t$, and A is normal, we have

$$AA^t = (PDV^t)(VDP^t) = (VDP^t)(PDV^t) = A^tA, \qquad \text{or}$$
$$PD^2P^t = VD^2V^t.$$

Thus, $(V^tP)D^2 = D^2(V^tP)$. Since the orthogonal matrix V^tP commutes with D^2, we have

$$(V^tP)_{i,j}(d_{i,i}^2 - d_{j,j}^2) = 0 \quad \text{for all } i \neq j.$$

If all $|d_{i,i}^2|$ are distinct, then V^tP is a diagonal, orthogonal matrix, and $V^tB = V^tAV = V^tPD$ is also diagonal. If the $|d_{i,i}|$ are not distinct, then V^tAV will be block diagonal, the blocks being multiples of orthogonal matrices.

Suppose now that $|d_{i,i}| = 0$, $i = r + 1 \cdots n$. The corresponding column norms of B are zero, and hence the column elements themselves must also be zero. We proceed as above except that we define P and D somewhat differently to avoid using D^{-1}. Define $D = \operatorname{diag}(d_{i,i}) = \operatorname{diag}(\|\mathbf{b}^1\| \cdots \|\mathbf{b}^r\|, 0 \cdots 0)$, and $P = (\mathbf{b}^1/d_{11} \cdots \mathbf{b}^r/d_{r,r}0 \cdots 0)$ so that again, $B = AV = PD$. Now $P^tP = I_{r,r}$ bounded by zero elements. This change, however, does not prevent the above proof from following exactly as before, and again we find that V^tP commutes with D^2. The corresponding conclusions follow.

We note in passing several points. If the pivot pairs are limited to $1 \leqq k \leqq k'$, $1 \leqq m \leqq n$, $(k < m)$, then we will obtain similar results for the k' largest singular values (or eigenvalues), provided $d_{k',k'}^2$ is distinct from $d_{k'+1,k'+1}^2$. To obtain the diagonal elements in increasing order, we need only minimize the S_k instead of maximizing them. The latter two remarks hold also for singular values of an $r \times n$ matrix, $r > n$.

Nash [8] mentions that the one-sided method may be used for symmetric matrices but that it may fail for singular matrices. We see no problem for singular matrices but rather for matrices having eigenvalues of equal absolute value. For example,

$$\begin{pmatrix} a & b & 0 & 0 \\ b & -a & 0 & 0 \\ 0 & 0 & -a & b \\ 0 & 0 & b & a \end{pmatrix} \text{ or } \begin{pmatrix} 0 & a & 0 & b \\ a & 0 & -b & 0 \\ 0 & -b & 0 & a \\ b & 0 & a & 0 \end{pmatrix}.$$

We also mention a set of *symmetric* matrices, which appear in chemical applications, which exhibit this property: these matrices are zero everywhere except on the subdiagonal and superdiagonal, and in the $(1, n)$ and $(n, 1)$ positions, where they have entries of one. When the above-described one-sided Jacobi algorithm is applied to these matrices, we obtain unitary blocks of larger sizes.

For one of these matrices, when n is 8, after 26 rotations, we obtain a 2×2 block:

$$\begin{pmatrix} 0.00 & 2.00 \\ 2.00 & 0.00 \end{pmatrix},$$

a 4×4 block:

$$\begin{pmatrix} 0.000000000 & 1.333649366 & 0.000000000 & 0.470509690 \\ 1.333649366 & 0.000000000 & -0.470509690 & 0.000000000 \\ 0.000000000 & -0.470509690 & 0.000000000 & 1.333649366 \\ 0.470509690 & 0.000000000 & 1.333649366 & 0.000000000 \end{pmatrix},$$

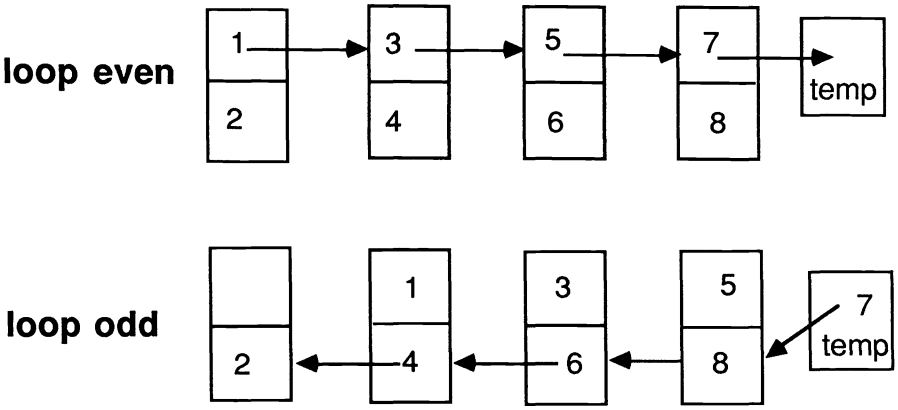
and two all zero rows. No problem occurs, although the matrix is clearly singular.

4. An algorithm for symmetric matrices. To implement a Jacobi procedure on the hypercube or a linear array of processors, the one-sided Jacobi algorithm offers many advantages over the block implementation. However, we wish to obtain convergence to a diagonal matrix rather than a block diagonal matrix. A simple change in the choice of angle makes this possible. We simply choose the classical Jacobi angle:

$$\tan 2\phi = 2\langle \mathbf{v}^k, \mathbf{a}^m \rangle / (\langle \mathbf{v}^k, \mathbf{a}^k \rangle - \langle \mathbf{v}^m, \mathbf{a}^m \rangle). \tag{5}$$

Note that $\tan 2\phi$ is expressed entirely in terms of the kth and mth vectors of A and of V. As Rutishauser [9] points out, it is possible to order the eigenvalues by using the larger angle of rotation; it is not clear whether or not this is a desirable thing to do. Convergence may be slowed somewhat by so ordering, because of the concomitant circulation of off-diagonal elements. On the other hand, for block Jacobi procedures, or when only the largest (or smallest) eigenvalues are required, ordering may be necessary.

We briefly discuss one possible implementation of one-sided Jacobi procedures on a hypercube. Suppose the columns of the matrix are distributed to the processors in pairs, and unit vectors made in the processors to accumulate the rotations. Multiple pairs of columns may also be distributed to the various processors depending on the sizes of the matrix and the cube. The parameters for the rotations depend only on the data resident in any processor; the columns may also be updated locally. After the $n/2$ rotations have been performed in parallel, the columns of both the matrix and the vector matrix are redistributed to neighboring processors. This may be done in a number of ways. See, for example, [1], [4] and [5]. One communication pattern [10] to derive appropriate rotations sets is illustrated for $n = 8$ below. For each sweep, a complete collection of rotation sets is obtained within a loop running from 0 to $n - 1$.

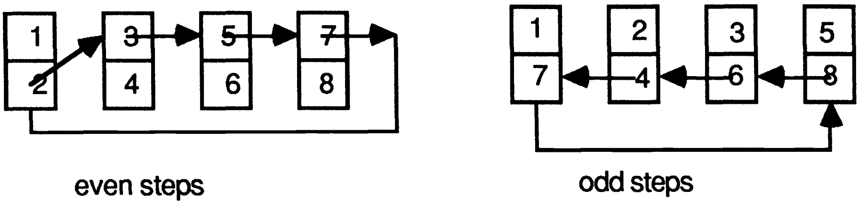


For $n = 8$ the rotations sets are:

(1,2)(3,4)(5,6)(7,8)
2(1,4)(3,6)(5,8)7
(2,4)(1,6)(3,8)(5,7)
4(2,6)(1,8)(3,7)5
(4,6)(2,8)(1,7)(3,5)
6(4,8)(2,7)(1,5)3
(6,8)(4,7)(2,5)(1,3)
8(6,7)(4,5)(2,3)1
(8,7)(6,5)(4,3)(2,1),

which is equivalent to the first set. This rotation pattern, which has been shown [5] to be equivalent to several others, has the advantage of having only one send and one receive; also it assumes no direct connection between the first and last processors. Furthermore, Luk and Park have proven convergence for this sequence of pivot pairs. How best to determine the termination criteria for the iteration is unclear. Either a fixed number of sweeps may be set, or global communication used to broadcast the size of the off-diagonal elements. We have chosen to allow node zero to determine when $|\tan \phi|$ is sufficiently small for all pivots during one complete sweep ($n(n - 2)/2$ rotations).

Another rotation pattern [4] is illustrated below; for this pattern convergence is not guaranteed. These rotation sets have the advantage that all processors may be active at every level and a sweep occurs after $n - 1$ rather than n steps. Again, the schema illustrated alternate:



Here the rotation sets are:

$$\begin{array}{l}
(1,2)(3,4)(5,6)(7,8)\\
(1,7)(2,4)(3,6)(5,8)\\
(1,4)(2,6)(3,8)(5,7)\\
(1,5)(4,6)(2,8)(3,7)\\
(1,6)(4,8)(2,7)(3,5)\\
(1,3)(6,8)(4,7)(2,5)\\
(1,8)(6,7)(4,5)(2,3)\\
(1,2)(8,7)(6,5)(4,3).
\end{array}$$

Some details for our implementation are available in [12], as are some initial timings and speed-ups.

Barry and Sameh [11] have reported that one-sided Jacobi (using the Hestenes angle) on an Alliant FX/8 outperformed "the most efficient" EISPACK routines for all orders considered, ($\leqq 400$) and LINPACK routines for the singular value problem 2–3 times. Since our one-sided Jacobi is of identical complexity to theirs (see the definition of the angles defined by (2) and (5)), we expect that the use of Jacobi algorithms on the hypercube, as well as other multiprocessor machines, will be competitive when carefully implemented.

REFERENCES

[1] R. P. BRENT AND F. T. LUK, *The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.

[2] B. CHARTRES, *Adaption of the Jacobi method for a computer with magnetic backing store*, Comput. J., 5 (1962), pp. 51–60.

[3] P. J. EBERLEIN, *On the use of hyperbolic and circular transformations in the scaling of matrices*, unpublished manuscript, 1965.

[4] ———, *Comments on some parallel Jacobi orderings*, Tech. Report, State University of New York, Buffalo, NY, 1986.

[5] F. T. LUK AND H. PARK, *On parallel Jacobi orderings*, EE-CEG-86-5, Cornell University, Ithaca, NY, 1986.

[6] M. R. HESTENES, *Inversion of matrices by biorthogonalization and related results*, SIAM J. Appl. Math., 6 (1958), pp. 51–90.

[7] H. F. KAISER, *The JK method: a procedure for finding the eigenvectors and eigenvalues of a real symmetric matrix*, Comput. J., 15 (1972), pp. 271–273.

[8] J. C. NASH, *A one-sided transformation method for the singular value decomposition and algebraic eigenproblem*, Comput. J., 18 (1977), pp. 74–76.

[9] H. RUTISHAUSER, *The Jacobi method for real symmetric matrices*, in Handbook for Automatic Computation, Vol. 2, Springer-Verlag, New York, Berlin, Heidelberg, 1971, pp. 202–211.

[10] R. A. WHITESIDE, N. S. OSTLUND AND P. G. HIBBARD, *A parallel Jacobi diagonalization algorithm for a loop multiple processor system*, IEEE Trans. Comput., C33 (1984), pp. 409–413.

[11] M. BARRY AND A. SAMEH, *Multiprocessor Jacobi algorithms for dense symmetric eigenvalue and singular value decompositions*, Proc. 86th Internat. Conference on Parallel Processing, IEEE Computer Society Press, August 1986.

[12] P. J. EBERLEIN, *On using the Jacobi method on the hypercube*, Proc. Second Hypercube Conference, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

SIAM J. Alg. Disc. Meth.
Vol. 8, No. 4, October 1987

023

AN EFFICIENT FACTORIZATION FOR THE GROUP INVERSE*

BERNARD F. LAMOND†

Abstract. An efficient algorithm is introduced for computing the group inverse of a square, singular matrix, in factorized form. The algorithm is based on the QR factorization with column pivoting and uses a technique of inversion by partitioning. The factorization is used to compute the group inverse solution of a singular system of equations. When only the solution vector is wanted, the group inverse does not need to be computed explicitly.

Key words. group inverse, matrix factorization, singular equations

AMS(MOS) subject classification. 65F05, 65F20

Introduction. For a singular matrix of index 1, the generalized inverse of Drazin (1958) is called the *group inverse*. As shown in Lamond and Puterman (1986), the solution of certain singular systems of equations can be expressed in terms of the group inverse of the matrix of coefficients. Such systems of equations are found, for example, in Markov and semi-Markov decision processes in operations research (see Veinott (1969) and Denardo (1971)). The numerical solution of these equations is required in the policy evaluation step of the Policy Iteration Method for finding an optimal policy.

In this paper, an algorithm of Wilkinson (1982) for obtaining a factorization of the Drazin inverse of a general matrix is specialized to the case of a matrix of index 1. This specialization gives the method of Robert (1968). By careful implementation of a formula for the inverse of a partitioned matrix, the algorithm is then modified to become nearly as efficient as the initial QR factorization of the matrix.

The modified algorithm, which is valid in general, is more efficient than the original one when the matrix is nearly of full rank and when the number of different right-hand vectors is small. These conditions are satisfied in the above applications. The solution computed with the above factorization is equal to the one defined by the group inverse of the matrix. The group inverse itself is never produced explicitly, however.

The paper is organized as follows. In Section 1, the singular systems of equations are defined and their solution is given in terms of the group inverse. In Section 2, the algorithm of Wilkinson (1982) is described in the special case when the group inverse exists. In Section 3, the modified algorithm is derived, to produce an efficient factorization of the group inverse. Finally, in Section 4 it is shown how the factorization can be used to compute the solution of the singular equations.

1. Statement of the problem.

Let A be an $n \times n$ matrix with real entries. The *index* of A, denoted $\operatorname{ind}(A)$, is the smallest nonnegative integer ℓ such that

$$\operatorname{rank}(A^{\ell+1}) = \operatorname{rank}(A^{\ell}).$$

The *Drazin generalized inverse* of A is the unique matrix A^D such that

$$AA^D = A^D A,$$

* Received by the editors May 23, 1986; accepted for publication July 20, 1987. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant A-5527.

† Department of Systems and Industrial Engineering, The University of Arizona, Tucson, Arizona 85721.

$$A^D A A^D = A^D,$$

$$A^D A^{\ell+1} = A^\ell,$$

where $\ell = \text{ind}(A)$ (see Drazin (1958) and Campbell and Meyer (1979)). When A is nonsingular, then $\text{ind}(A) = 0$ and $A^D = A^{-1}$, the ordinary inverse of A. We are interested in the case when A is singular, but such that $\text{ind}(A) = 1$. In this special case, we write $A^D = A^\#$ and we say that $A^\#$ is the *group inverse* of A.

Now suppose b is an n-vector such that the singular system of equations

$$Ax = b \tag{1.1}$$

has a solution. Then b is said to be *compatible* with A and there are infinitely many vectors x satisfying (1.1). One celebrated solution is the vector $\hat{x}$ such that

$$\hat{x} = A^\dagger b, \tag{1.2}$$

where $A^\dagger$ is the Moore-Penrose pseudo-inverse of A. Of all the vectors x satisfying (1.1), $\hat{x}$ has the smallest ℓ_2-norm. Stable and efficient algorithms for computing $\hat{x}$ are well understood (see, e.g., Golub and Van Loan (1983)), and the corresponding computer software is widely available.

There are situations, however, for which $\hat{x}$ is not the most suitable solution. For example, suppose $b_1, b_2, \cdots, b_k$ are given vectors and that we are interested in finding some vectors $x_1, x_2, \cdots, x_k$ such that the following system of equations is satisfied:

$$\begin{pmatrix} A & 0 & \cdots & \cdots & 0 \\ I & A & \ddots & & \vdots \\ 0 & I & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I & A \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_k \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{pmatrix}, \tag{1.3}$$

where A is a singular $n \times n$ matrix such that $\text{ind}(A) = 1$. Such equations must be solved for finding an optimal policy in a Markov decision process when the interest rate is small (see Veinott (1969) and Lamond and Puterman (1986)). The solution to (1.3) is completely characterized by the group inverse of the coefficient matrix A, according to the following theorem from Lamond and Puterman (1986).

THEOREM 1.1. *Suppose* $\text{ind}(A) = 1$ *and* b_1 *is compatible with* A. *Then the system of equations* (1.3) *has a solution. Moreover, with* $x_0 = 0$ *and* b_{k+1} *arbitrary, we have*

$$x_i = A^\#(b_i - x_{i-1}) + W b_{i+1}, \tag{1.4}$$

for $i = 1, \cdots, k$, *where*

$$W = I - AA^\#. \tag{1.5}$$

The matrix W is called the *eigenprojection* of A (see Rothblum (1981)), and it is directly verified that $AW = WA = 0$. Hence Theorem 1.1 is a simple consequence of the fact that a vector x is a solution of equation (1.1) if and only if

$$x = A^\# b + Wy$$

for some (arbitrary) vector y (see Lamond and Puterman (1986)).

The solution of equation (1.3) can also be expressed in terms of the Moore-Penrose inverse $A^\dagger$, but the formula is more awkward than equation (1.4). Let $\tilde{W} = I - AA^\dagger$. It is straightforward to show that the solution to (1.3) is

$$x_i = (I - \tilde{W})A^\dagger(b_i - x_{i-1}) + \tilde{W}b_{i+1}. \tag{1.6}$$

In general, the term $\tilde{W}A^\dagger$ is not equal to zero. The consequence is that (1.6) does not lead to an expression for x_i in terms of powers of $A^\dagger$. Such an expression is derived in Lamond and Puterman (1986) using the powers of $A^\#$, based on equation (1.4).

The object of this paper is to propose an efficient algorithm to obtain $A^\#$ in factorized form, and to use the factorization to compute the solution vectors $x_1, \cdots, x_k$ of equation (1.3). The algorithm is particularly efficient when $m << n$ and $k << n$, where A is $n \times n$, $\text{rank}(A) = n - m$ and k is the number of blocks in (1.3). The factorization can also be used to compute $A^\#$ explicitly, but this step can be skipped when only the solution vectors are needed.

Equations of the form (1.3) have to be solved in the Policy Evaluation step of the Policy Improvement method for finding an optimal policy in Markov Decision Processes (see Veinott (1969) and Lamond and Puterman (1986)). The algorithm can also be used to solve the more complicated systems encountered in Semi-Markov Decision Processes (see Denardo (1971) and Lamond (1985)), in which the subdiagonal blocks of equation (1.3) are nonzero.

2. Factorization of the group inverse. The factorization algorithm is based on the well known fact (see, e.g., Campbell and Meyer (1979)) that for an $n \times n$ matrix A with $\text{ind}(A) = 1$, there exists a pair of nonsingular matrices S and B such that

$$A = S \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} S^{-1}, \tag{2.1}$$

where S is $n \times n$ and B is $(n - m) \times (n - m)$, with $\text{rank}(A) = n - m$. The group inverse $A^\#$ is then given by

$$A^\# = S \begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix} S^{-1}. \tag{2.2}$$

The algorithm of Wilkinson (1982), which is valid also when $\text{ind}(A) > 1$, uses standard reduction methods to produce such a transformation in which S, S^{-1} and B^{-1} are kept in factorized form. In Section 3, we will see how the matrix B can also be kept in factorized form, leading to a substantial reduction in the amount of arithmetic required, assuming $m << n$. Keeping this in mind, the former method will be referred to as the *slow algorithm*, and the latter will be called the *fast algorithm*.

The following results are well known (see Robert (1968) and Campbell and Meyer (1979, Lemma 7.7.2, Theorem 7.7.8)). They provide the mathematical basis from which the computational algorithms are derived.

LEMMA 2.1. *Suppose A is an $n \times n$ matrix such that* $\text{rank}(A) = n - m$, *with $1 \leq m < n$, and*

$$A = S \begin{pmatrix} B_{11} & B_{12} \\ 0 & 0 \end{pmatrix} S^{-1} \tag{2.3}$$

for some nonsingular matrix S, where B_{11} is an $(n-m) \times (n-m)$ matrix. Then B_{11} is nonsingular if and only if $\text{ind}(A) = 1$.

THEOREM 2.2. *Suppose A is an $n \times n$ matrix satisfying equation* (2.3) *with B_{11} nonsingular. Then*

$$A^{\#} = S \begin{pmatrix} B_{11}^{-1} & B_{11}^{-1}B_{11}^{-1}B_{12} \\ 0 & 0 \end{pmatrix} S^{-1} \tag{2.4}$$

and

$$W = I - AA^{\#} = S \begin{pmatrix} 0 & -B_{11}^{-1}B_{12} \\ 0 & I \end{pmatrix} S^{-1}. \tag{2.5}$$

To decompose the matrix A as in equation (2.3), two steps are required. In the first step, the matrix is reduced to a simpler form (e.g., upper triangular) using standard row transformations. In the second, the inverse transformations are applied to the columns of the reduced matrix, giving (2.3). This second step is necessary because (2.3) defines a similarity transformation.

One method that is both numerically stable and computationally efficient for the reduction of a singular matrix is the QR factorization with column pivoting (see Golub and Van Loan (1983)). It decomposes A as

$$A = QRT, \tag{2.6}$$

where Q is an orthogonal matrix (computed in factorized form), T is a permutation matrix and

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}, \tag{2.7}$$

with R_{11} a nonsingular $(n-m) \times (n-m)$ upper triangular matrix. The number of arithmetic operations is dominated by $\frac{2}{3}(n-m)^3$ flops (here a *flop*, or *floating-point operation*, is defined as one multiplication plus one addition; see Golub and Van Loan (1983, p. 32)).

Other reduction methods can be used instead of the QR factorization. For example, in the special case when $A = I - P$, where P is a stochastic matrix, the Gaussian elimination procedure is known to be numerically stable (see Harrod and Plemmons (1984) and Funderlic and Plemmons (1981)). Further, in this case $\operatorname{rank}(A) = n - m$ is known in advance because m is the number of recurrent classes of P. Since Gaussian elimination requires only $\frac{1}{3}(n-m)^3$ flops, it is more economical than the QR factorization. See Lamond (1985) for a description of the Gaussian elimination version of the group inverse algorithms. Also, a more direct method that takes full advantage of the stochastic matrix structure, but without using the group inverse theory, is given in Veinott (1969).

On the other hand, if the matrix A is expected to be ill-conditioned, then both of the above reduction methods can be numerically unstable. In this case, the preferred reduction method is the singular value decomposition (see, e.g., Golub and Van Loan (1983)). While it requires a larger amount of arithmetic for the reduction, the singular value decomposition has guaranteed stability. It is a straightforward exercise to modify the algorithms of this paper to use the singular value decomposition instead of the QR factorization.

Now to convert equation (2.6) into the required form as in equation (2.3), observe that

$$A = QRT = Q(RTQ)Q^{-1}, \tag{2.8}$$

with

$$RTQ = B = \begin{pmatrix} B_{11} & B_{12} \\ 0 & 0 \end{pmatrix} \tag{2.9}$$

as required, by (2.7). The *slow group inverse factorization* algorithm is given in Figure 2.1. The amount of arithmetic is $\frac{2}{3}(n-m)^3$ at Step 1, $n^2(n-m)$ at Step 2 and $\frac{1}{3}(n-m)^3$ at Step 3, for a total of $2n^3$ flops (when $m << n$). This algorithm is a specialization of the method of Wilkinson (1982). If singular value decomposition was used for the reduction, this algorithm would be a specialization of Algorithm 12.5.1 of Campbell and Meyer (1979).

```
Step 1.  Compute $Q$, $R$ and $T$ such that $A = QRT$
         (Using QR factorization with column pivoting).
Step 2.  Compute $B = RTQ$
         (Applying $T$ and $Q$ to the nonzero rows of $R$).
Step 3.  Compute $P$, $L$ and $U$ such that $B_{11} = PLU$
         (Using Gaussian elimination).
Step 4.  If $B_{11}$ is singular then "Error: ind($A$) > 1"
         else stop (we have all the necessary information).
```

FIG. 2.1. *The slow group inverse factorization.*

Note that at Step 3, ordinary Gaussian elimination is used (because B_{11} is nonsingular), to produce a permutation matrix P, a unit lower triangular matrix L (in factorized form) and a nonsingular upper triangular matrix U, all of dimension $(n-m) \times (n-m)$. The error exit, at Step 4, is provided by Lemma 2.1.

To illustrate the process, let us consider a numerical example with $n = 3$ and $m = 1$. Let

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}.$$

Then the group inverse of A exists and is given by

$$A^{\#} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & 2 \\ -1 & 1 & -1 \end{pmatrix}.$$

Applying the QR algorithm to A, we obtain that $A = QRT$, where

$$Q = \begin{pmatrix} -\sqrt{2}/2 & 0 & \sqrt{2}/2 \\ 0 & 1 & 0 \\ \sqrt{2}/2 & 0 & \sqrt{2}/2 \end{pmatrix},$$

$$R = \begin{pmatrix} -\sqrt{2} & 0 & \sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Hence

$$R_{11} = \begin{pmatrix} -\sqrt{2} & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad R_{12} = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}.$$

From this, we get $B = RTQ$, so that

$$B_{11} = \begin{pmatrix} 1 & \sqrt{2} \\ \sqrt{2}/2 & 0 \end{pmatrix} \quad \text{and} \quad B_{12} = \begin{pmatrix} -1 \\ \sqrt{2}/2 \end{pmatrix}.$$

Finally, Gaussian elimination yields $B_{11} = PLU$, with

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 \\ \sqrt{2}/2 & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} 1 & \sqrt{2} \\ 0 & -1 \end{pmatrix}.$$

From this, it is straightforward to verify that

$$B_{11}^{-1} = U^{-1}L^{-1}P^{-1} = \begin{pmatrix} 0 & \sqrt{2} \\ \sqrt{2}/2 & -1 \end{pmatrix}.$$

Substituting into equation (2.4) with $S = Q$ and $S^{-1} = Q'$, we obtain $A^{\#}$ as given above.

Observe that while the algorithm of Figure 2.1 produces all the building blocks for computing the group inverse $A^{\#}$, it is not necessary to produce the matrix $A^{\#}$ explicitly. The algorithms of Section 4 will compute the solution vectors directly, without computing the group inverse itself. This approach is analogous to the method of forward and back substitution for solving nonsingular equations using the LU factorization of the matrix of coefficients.

3. Factorization of the intermediate matrix. In this section, we introduce a device by which the amount of arithmetic required at Steps 2 and 3 of the algorithm of Figure 2.1 is reduced by an order of magnitude, when $m << n$. Hence the *fast algorithm* so obtained requires a total of about $\frac{2}{3}n^3$ flops, which is basically the amount of work performed in the initial QR factorization, at Step 1. This device is in fact a careful implementation of a standard formula for inverting a matrix by partitioning. The following lemma is derived in Faddeeva (1959).

LEMMA 3.1. *Let X be a nonsingular $n \times n$ matrix and Y its inverse. Suppose that X and Y are partitioned such that*

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \quad \textit{and} \quad Y = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix},$$

where X_{11} is nonsingular. Then Y_{22} is also nonsingular and

$$X_{11}^{-1} = Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}. \tag{3.1}$$

Of course, the idea here is to identify the above submatrix X_{11} with the $(n - m) \times (n - m)$ submatrix B_{11} of equation (2.3). A factorization for the matrix Y will be obtained such that Y_{22} can be produced in $O(mn^2)$ flops. The reduction of Y_{22}, using Gaussian elimination, then takes only $\frac{1}{3}m^3$ flops. Both counts are negligible compared to $\frac{2}{3}n^3$.

This can be done as follows. Consider the upper triangular matrix R of equation (2.7), which is produced at Step 1 of the algorithm of Figure 2.1. Now define a nonsingular matrix D such that

$$D = \begin{pmatrix} R_{11} & R_{12} \\ 0 & I \end{pmatrix}, \tag{3.2}$$

where I is the $m \times m$ identity matrix and R_{11} and R_{12} are as in (2.7). The inverse of D can be written immediately as

$$D^{-1} = \begin{pmatrix} R_{11}^{-1} & -R_{11}^{-1}R_{12} \\ 0 & I \end{pmatrix}, \tag{3.3}$$

in which R_{11}^{-1} itself does not need to be computed since R_{11} is upper triangular, so that the product by R_{11}^{-1} can be done by back substitution.

Now recall that $B = RTQ$ at Step 2 of the algorithm of Figure 2.1. Replacing R by D, we have the desired matrix $X = DTQ$, such that

$$X = \begin{pmatrix} B_{11} & B_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{3.4}$$

where $C = TQ$ is partitioned as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}.$$

The matrix C is of no interest in itself. The important result here is that the submatrix B_{11} of equation (2.9) is the same as that of equation (3.4). Further, the matrix $Y = X^{-1}$ is directly available in factorized form, because

$$Y = X^{-1} = Q^{-1}T^{-1}D^{-1}, \tag{3.5}$$

where $Q^{-1} = Q'$ is already factorized, T^{-1} is just a permutation matrix and D^{-1} is given in (3.3).

THEOREM 3.2. *Both B_{11} and Y_{22} are nonsingular if and only if* $\operatorname{ind}(A) = 1$. *Moreover,*

$$B_{11}^{-1} = Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}.$$

Proof. The proof is immediate from Lemmas 2.1 and 3.1. □

Now if we actually compute the matrix Y directly from equation (3.5), we need $\frac{1}{2}(n-m)^2 n$ flops to compute D^{-1} plus n^3 flops to do the product by T^{-1} and Q^{-1}. This is more work than for both Steps 2 and 3 of the slow algorithm. The key idea here is that we need only to compute the submatrix Y_{22}, which can be done as follows:

$$\begin{pmatrix} Y_{12} \\ Y_{22} \end{pmatrix} = Q^{-1}T^{-1} \begin{pmatrix} -R_{11}^{-1}R_{12} \\ I \end{pmatrix}, \tag{3.6}$$

where we used equation (3.3). The amount of work for computing Y_{22} by equation (3.6) is then $\frac{1}{2}m(n-m)^2$ flops for $R_{11}^{-1}R_{12}$, and mn^2 flops for the product by $Q^{-1}T^{-1}$. This is negligible when $m << n$, by comparison to Step 1.

The *fast group inverse factorization* algorithm is given in Figure 3.1. Again, at Step 3, ordinary Gaussian elimination is used, but now to factorize the $m \times m$ matrix Y_{22} into a permutation matrix P, a unit lower triangular matrix L and an upper triangular matrix U. In addition to being computationally efficient in terms of the amount of arithmetic required, the fast algorithm is also very economical in its storage requirements. In fact, it is possible to organize the computations in such a way that the whole factorization overwrites the $n \times n$ array containing the original matrix A.

Step 1. Compute Q, R and T such that $A = QRT$
(using QR factorization with column pivoting).
Step 2. Compute Y_{22} as in equation (3.6).
Step 3. Compute P, L and U such that $Y_{22} = PLU$
(using Gaussian elimination).
Step 4. If Y_{22} is singular then "Error: $\operatorname{ind}(A) > 1$"
else stop (we have all the necessary information).

FIG. 3.1. *The fast group inverse factorization.*

Let us apply the fast algorithm to the 3×3 example of Section 2. Step 1 is the same as for the slow algorithm. At Step 2, however, we have

$$R_{11}^{-1}R_{12} = \begin{pmatrix} -\sqrt{2}/2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix},$$

so that

$$\begin{pmatrix} -R_{11}^{-1}R_{12} \\ I \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

and

$$\begin{pmatrix} Y_{12} \\ Y_{22} \end{pmatrix} = Q^{-1}T^{-1} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sqrt{2}/2 \\ 1 \\ \sqrt{2}/2 \end{pmatrix}.$$

Hence $Y_{22} = (\sqrt{2}/2)$ so that its PLU factorization, at Step 3, is simply $P = L = (1)$ and $U = (\sqrt{2}/2)$.

Now let us verify that Theorem 3.2 gives the same matrix B_{11}^{-1} as we found in Section 2. We have $Y_{22}^{-1} = (\sqrt{2})$. Also, equations (3.3) and (3.5) give

$$\begin{pmatrix} Y_{11} \\ Y_{21} \end{pmatrix} = Q^{-1}T^{-1} \begin{pmatrix} R_{11}^{-1} \\ 0 \end{pmatrix} = \begin{pmatrix} 1/2 & \sqrt{2}/2 \\ 0 & 0 \\ -1/2 & \sqrt{2}/2 \end{pmatrix},$$

so that

$$Y_{11} = \begin{pmatrix} 1/2 & \sqrt{2}/2 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad Y_{21} = (\,-1/2 \quad \sqrt{2}/2\,).$$

Applying Theorem 3.2, we have

$$\begin{aligned} B_{11}^{-1} &= \begin{pmatrix} 1/2 & \sqrt{2}/2 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} -\sqrt{2}/2 \\ 1 \end{pmatrix} (\sqrt{2})(\,-1/2 \quad \sqrt{2}/2\,) \\ &= \begin{pmatrix} 0 & \sqrt{2} \\ \sqrt{2}/2 & -1 \end{pmatrix} \end{aligned}$$

which is the correct value.

4. Solution of singular systems of equations.

In this section, we show how the factorization of the previous sections can be used to compute the solution of equation (1.3), based on Theorem 1.1. Suppose A is decomposed as in equation (2.3), so that $A^{\#}$ and W are given by (2.4) and (2.5), respectively. Define

$$y_i = S^{-1}x_i = \begin{pmatrix} y_{1,i} \\ y_{2,i} \end{pmatrix} \quad \text{for} \quad i = 0, \cdots, k,$$

and

$$c_i = S^{-1}b_i = \begin{pmatrix} c_{1,i} \\ c_{2,i} \end{pmatrix} \quad \text{for} \quad i = 1, \cdots, k+1.$$

THEOREM 4.1. *Suppose b_1 is compatible with A. Then*

$$y_{1,i} = B_{11}^{-1}(c_{1,i} - y_{1,i-1} - B_{12}c_{2,i+1}), \tag{4.1}$$

$$y_{2,i} = c_{2,i+1} \tag{4.2}$$

for $i = 1, \cdots, k$.

Proof. By induction on i. The result is trivially true for $i = 0$ if we define $c_{1,0} = 0$ and $y_{1,-1} = 0$, because $x_0 = 0$ and $c_{2,1} = 0$ since b_1 is compatible. Now suppose $y_{2,i-1} = c_{2,i}$. Then equation (1.4) gives

$$\begin{pmatrix} y_{1,i} \\ y_{2,i} \end{pmatrix} = \begin{pmatrix} B_{11}^{-1} & B_{11}^{-1}B_{11}^{-1}B_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c_{1,i} - y_{1,i-1} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & -B_{11}^{-1}B_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} c_{1,i+1} \\ c_{2,i+1} \end{pmatrix}$$

and the result follows. □

If the matrix A was factorized using the slow algorithm, the solution vectors $x_1, \cdots, x_k$ can be computed as in Figure 4.1. The products by Q and Q^{-1}, at lines 1, 3 and 7 are done using the factorization for Q, and take n^2 flops each. The product by B_{12}, at line 4, takes $m(n-m)$ flops and the solution $y_{1,i}$ at line 5 is obtained by forward and back substitution, in $(n-m)^2$ flops. Hence the total work required by the *slow* solution algorithm is $(3k+1)n^2$ when $m << n$. This is faster than the *fast* solution algorithm that will be derived below. But with $k << n$, the amount of arithmetic required for the solution is negligible by comparison to that required in the factorization. Hence the labels *slow* and *fast* refer to the factorization employed.

```
1. Compute $c_1 = Q^{-1}b_1$ and let $y_0 = 0$.
2. For $i = 1, \cdots, k$ do:
3.     Compute $c_{i+1} = Q^{-1}b_{i+1}$;
4.     Compute $v_1 = c_{1,i} - y_{1,i-1} - B_{12}c_{2,i+1}$;
5.     Solve $PLUy_{1,i} = v_1$;
6.     Set $y_{2,i} = c_{2,i+1}$;
7.     Compute $x_i = Qy_i$.
8. Stop.
```

FIG. 4.1. *The slow solution algorithm.*

To illustrate the solution process, consider again the matrix A of Section 2 with $n = 3$ and $m = 1$, and suppose we want to solve equation (1.3) with $k = 2$,

$$b_1 = \begin{pmatrix} 3 \\ 2 \\ -3 \end{pmatrix} \quad \text{and} \quad b_2 = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}.$$

Since

$$W = I - AA^{\#} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

Theorem 1.1 gives the solution

$$x_1 = \begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix},$$

where we chose $b_3 = 0$ arbitrarily.

Following the algorithm of Figure 4.1, we have, at lines 1 and 3,

$$c_1 = Q^{-1}\begin{pmatrix} 3 \\ 2 \\ -3 \end{pmatrix} = \begin{pmatrix} -3\sqrt{2} \\ 2 \\ 0 \end{pmatrix} \quad \text{and} \quad c_2 = Q^{-1}\begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ 3 \\ 3\sqrt{2} \end{pmatrix}.$$

Then at lines 4, 5 and 6, we have

$$v_1 = \begin{pmatrix} -3\sqrt{2} \\ 2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ \sqrt{2}/2 \end{pmatrix} (3\sqrt{2}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

$$y_{1,1} = U^{-1}L^{-1}P^{-1} \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -\sqrt{2} \\ 1 \end{pmatrix}$$

and $y_{2,1} = (3\sqrt{2})$. Hence line 7 gives

$$x_1 = Q \begin{pmatrix} -\sqrt{2} \\ 1 \\ 3\sqrt{2} \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix},$$

which is the correct value. Similarly, the second iteration gives $c_3 = 0$,

$$v_1 = \begin{pmatrix} 2\sqrt{2} \\ 2 \end{pmatrix}, \quad y_{1,2} = \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix}, \quad y_{2,2} = (0)$$

and

$$x_2 = Q \begin{pmatrix} 2\sqrt{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix}$$

as required.

Suppose now that the matrix A was factorized using the fast algorithm. Then the solution algorithm must be organized carefully since the intermediate matrices are in factorized form. To get the product by B_{12}, we proceed as follows. Suppose $B_{12}z_2$ is required for some m-vector z_2. Then by equations (2.7) and (2.9), we have

$$\begin{aligned} B_{12}z_2 &= (\, B_{11} \quad B_{12}\,) \begin{pmatrix} 0 \\ z_2 \end{pmatrix} \\ &= (\, R_{11} \quad R_{12}\,) TQ \begin{pmatrix} 0 \\ z_2 \end{pmatrix}. \end{aligned} \tag{4.3}$$

This product can be done in at most $2n^2$ flops.

To get the product by B_{11}^{-1}, we proceed as follows. Suppose $B_{11}^{-1}z_1$ is required for some $(n-m)$-vector z_1. Then by Theorem 3.2, we have

$$B_{11}^{-1}z_1 = (Y_{11}z_1) - Y_{12}Y_{22}^{-1}(Y_{21}z_1).$$

By equations (3.3) and (3.5), the products $Y_{11}z_1$ and $Y_{21}z_1$ can be obtained simultaneously with

$$\begin{pmatrix} Y_{11} \\ Y_{21} \end{pmatrix} z_1 = Q^{-1}T^{-1} \begin{pmatrix} R_{11}^{-1}z_1 \\ 0 \end{pmatrix}. \tag{4.4}$$

This can be done in at most $\frac{3}{2}n^2$ flops.

Further, the product by Y_{12} can be done as follows. If $Y_{12}z_2$ is required for some m-vector z_2, we can use

$$\begin{pmatrix} Y_{12} \\ Y_{22} \end{pmatrix} z_2 = Q^{-1}T^{-1} \begin{pmatrix} -R_{11}^{-1}R_{12}z_2 \\ z_2 \end{pmatrix}, \tag{4.5}$$

which can be done with at most $\frac{3}{2}n^2$ flops (with $m << n$).

```
1.  Compute $c_1 = Q^{-1}b_1$ and let $y_0 = 0$.
2.  For $i = 1, \cdots, k$ do:
3.      Compute $c_{i+1} = Q^{-1}b_{i+1}$;
4.      Compute $w_1 = B_{12}c_{2,i+1}$ as in equation (4.3);
5.      Compute $v_1 = c_{1,i} - y_{1,i-1} - w_1$;
6.      Compute $w_1 = Y_{11}v_1$ and $w_2 = Y_{21}v_1$ as in equation (4.4);
7.      Solve $PLUz_2 = w_2$;
8.      Compute $z_1 = Y_{12}z_2$ as in equation (4.5);
9.      Compute $y_{1,i} = w_1 - z_1$;
10.     Set $y_{2,i} = c_{2,i+1}$;
11.     Compute $x_i = Qy_i$.
12. Stop.
```

FIG. 4.2. *The fast solution algorithm.*

Of course, the product by Y_{22}^{-1} can be done using the PLU factorization of Y_{22}. The fast solution algorithm is given in Figure 4.2. It was obtained from the slow algorithm of Figure 4.1 by replacing line 4 by lines 4 and 5 and line 5 by lines 6 to 9. The amount of work is $(7k+1)n^2$ flops, which is negligible by comparison to that of the factorization, when $k << n$.

Let us return to the numerical example. In the first iteration, lines 4 and 5 give

$$w_1 = (\,R_{11} \quad R_{12}\,)TQ\begin{pmatrix} 0 \\ 0 \\ 3\sqrt{2} \end{pmatrix}$$

$$= \begin{pmatrix} -\sqrt{2} & 0 & \sqrt{2} \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix} = \begin{pmatrix} -3\sqrt{2} \\ 3 \end{pmatrix}$$

and

$$v_1 = \begin{pmatrix} -3\sqrt{2} \\ 2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -3\sqrt{2} \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

which is what we had at line 4 of the slow solution algorithm. Then at lines 6 to 10 of Figure 4.2, we have

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = Q^{-1}T^{-1}\begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} = \begin{pmatrix} -\sqrt{2}/2 \\ 0 \\ -\sqrt{2}/2 \end{pmatrix},$$

$$z_2 = U^{-1}L^{-1}P^{-1}\,(\,-\sqrt{2}/2\,) = (\,\sqrt{2}\,)\,(\,-\sqrt{2}/2\,) = (\,-1\,),$$

$$\begin{pmatrix} z_1 \\ \cdots \end{pmatrix} = Q^{-1}T^{-1}\begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} \sqrt{2}/2 \\ -1 \\ -\sqrt{2}/2 \end{pmatrix},$$

$$y_{1,1} = \begin{pmatrix} -\sqrt{2}/2 \\ 0 \end{pmatrix} - \begin{pmatrix} \sqrt{2}/2 \\ -1 \end{pmatrix} = \begin{pmatrix} -\sqrt{2} \\ 1 \end{pmatrix}$$

and $y_{2,1} = (3\sqrt{2})$. This is precisely the same solution as with the slow algorithm. The second iteration works similarly.

Acknowledgments. I would like to thank Professors R.L. Johnston, Uri Ascher and Peter Kiernan for their comments and encouragements at various stages of this work.

REFERENCES

S.L. CAMPBELL AND C.D. MEYER, JR. (1979), *Generalized Inverses of Linear Transformations,* Pitman, London.

E.V. DENARDO (1971), *Markov renewal programs with small interest rates,* Ann. Math. Statist., 42, pp. 477–496.

M.P. DRAZIN (1958), *Pseudo-inverses in associative rings and semigroups,* Amer. Math. Monthly, 65, pp. 506–514.

V.N. FADDEEVA (1959), *Computational Methods of Linear Algebra,* Dover, New York.

R.E. FUNDERLIC AND R.J. PLEMMONS (1981), LU *decomposition of* M-*matrices by elimination without pivoting,* Linear Algebra Appl., 41, pp. 99–110.

G.H. GOLUB AND C.F. VAN LOAN (1983), *Matrix Computations,* The Johns Hopkins University Press, Baltimore, MD.

W.J. HARROD AND R.J. PLEMMONS (1984), *Comparison of some direct methods for computing stationary distributions of Markov chains,* SIAM J. Sci. Statist. Comput., 5, pp. 453–469.

B.F. LAMOND (1985), *Matrix methods in queueing and dynamic programming,* Ph.D. dissertation, The University of British Columbia, Vancouver, Canada.

B.F. LAMOND AND M.L. PUTERMAN (1986), *Generalized inverses in discrete time Markov decision processes,* Submitted for publication.

P. ROBERT (1968), *On the group-inverse of a linear transformation,* J. Math. Anal. Appl., 22, pp. 658–669.

U.G. ROTHBLUM (1981), *Resolvent expansions of matrices and applications,* Linear Algebra Appl., 38, pp. 33–49.

A.F. VEINOTT, JR. (1969), *Discrete dynamic programming with sensitive discount optimality criteria,* Ann. Math. Statist., 40, pp. 1635–1660.

J.H. WILKINSON (1982), *Note on the practical significance of the Drazin inverse,* in Recent Applications of Generalized Inverses, S.L. Campbell, ed., Pitman, Boston., pp. 82–99.